International Journal of Imaging Systems and Technology (IJIST), vol. 17, no. 3, pp. 174-189, 2007. (Pre-print version)

Color aided motion-segmentation and object tracking for video sequences semantic analysis

Alexia Briassouli, Vasileios Mezaris, Ioannis Kompatsiaris

Informatics and Telematics Institute
Centre for Research and Technology Hellas
Thermi-Thessaloniki, 57001, Greece
{abria,bmezaris,ikom}@iti.gr

ABSTRACT

The high rates at which digital multimedia is being generated and used makes it necessary to develop systems that can process it in an efficient manner. This can be achieved by extracting semantics from processing the video's low-level information. We present a novel algorithm which fuses color and motion information, in order to extract semantics from the video sequence. The motion estimates are processed statistically to give areas of activity in the video. Color segmentation is applied to these areas, and also to their complementary regions in each frame, in order to achieve the moving object segmentation. The extracted color layers in the activity and background areas are compared using the Earth Mover's Distance (EMD), and a novel method, which we introduce, and which is based on a likelihood ratio test (LRT). The segmentation results of our LRT-based approach are shown to be more robust than the EMD results, and both methods are shown to be more accurate than existing combined color-motion approaches. Furthermore, the LRT method allows the retrieval of additional semantics, namely of "maps" that indicate with what likelihood a pixel belongs to a moving object. The areas of activity can be used to retrieve semantics for the kind of activity taking place. The color-aided segmentation of the moving entities provides a full description of their appearance, so it can be used, for example, to classify the video based on the objects in it. Experiments with real sequences show that this method leads to accurate results and useful semantics.

I. INTRODUCTION

The extraction of semantics from video has attracted significant attention lately, as the increasingly vast amounts of multimedia data that are becoming available are making it very difficult to access and manipulate it via traditional methods. The results of image and video processing methods, such as motion estimation, feature extraction, and object segmentation, can be incorporated into systems that can retrieve semantics from them. These higher level concepts (semantics) can be used to extract the main content from the video, and can consequently help in its more efficient analysis, interpretation and further processing.

In certain applications, such as sports, rules of the games have been used to extract semantics from the motion and appearance features of a video [1], [2]. For example, in sports video semantic analysis [3], the dominant color of the frames helps determine the regions where the event is taking place, such as the football field or the tennis court. The extraction of additional appearance features, such as the location of the lines in a field or a net in a tennis court, also contribute to the extraction of semantics [4]. Motion information analysis plays a fundamental role in understanding the events taking place in a video, as it shows what kind of action is occurring, and how it evolves with time. Sports analysis systems often combine color, texture and motion features in order to analyze the video content. For example, the trajectories of objects and their interactions have been used to detect events in soccer [5], [6]. However, these methods use manually obtained trajectories, and could fail if motion information extracted from the video processing was used, since it is likely to be noisy. Other motion based methods [7] rely heavily on rules of the game, or prior knowledge about the camera setup. Although these approaches give satisfactory results for sports analysis purposes, they suffer from the drawback of being too application-oriented. In order to achieve satisfactory player segmentation results and activity classification, they rely on rules concerning the specific application (sport) and appearance characteristics of the corresponding environment (soccer field, tennis court etc). The extraction of semantics from video sequences is also important for surveillance applications, where it is often of interest to detect “abnormal” events or activities [8], [9]. Commonly used video motion detection (VMD) surveillance systems are based only on velocity estimation, and thus often suffer from false alarms, caused by nonmotion events (e.g. changes in scene illumination). This has led to the development of surveillance systems that can characterize the detected motion in a video, in relation to its context, such as scene appearance, or motion history. In these applications, the motion trajectories are used to train the system to detect unusual events [10], [11]. These approaches are also application-dependent, as they often use prior knowledge about the scene structure in order to give meaningful results, and are trained to expect specific types of events. Finally, when video processing methods use only the motion information to extract semantics, they have been shown to be sensitive to occlusions and inaccuracies in the velocity estimates [10].

Various approaches have been developed for the extraction of moving objects from a video sequence. In [12], a very effective background removal method is presented, which is also able to deal with varying illumination variations, small background motions and camera displacements. However, this method is based only on color, and can extract color homogeneous foreground regions, which introduce the need for “heuristics” for the detection of moving persons (a person is formed when two homogeneous color blobs are located one above the other). This limits the applicability of that approach, which would fail when applied to videos of other kinds of motion (not only humans moving), as in our work. Several methods have been developed in the literature for the combination of motion and color, in order to reliably segment moving objects. In [13], [14], [15], color clusters are formed in a video, and are then compared in order to detect motion between regions of similar color. Regions with similar color and location [13], [14], or motion [15], are clustered together, as they are considered to belong to the same moving object. These methods lead to good segmentation results when the color contrast between static and moving areas is high, but can encounter difficulties if the initial color segmentation is not accurate, as they will be searching for motion in incorrectly segmented frame regions. The method of [13] explicitly focuses on the case of rigid object motions, as it is assumed that image motion across a color-homogeneous cluster can be approximated by the motion of its centroid. Nonrigid motion is analyzed by focusing on the tracking of homogeneous color-blobs [14]. However these methods are still expected to encounter problems, such as false alarms, introduced by oversegmenting the video sequence. Also, they are based on clustering pixels with simultaneously similar color and location or motion [14], [15], so they cannot deal with differently colored objects, with differently moving parts, which may appear in very different locations in each frame (e.g. the leg and arm of a tennis player).

In this paper, we propose a system that can be used for the extraction of video semantics in more general applications, by integrating motion and color information. Contrary to existing methods, it does not limit its use of motion to the object trajectories, but extracts areas of activity by accumulating motion information over all frame pixels, and over several frames. For the case of a moving camera, its motion is first compensated for in a global motion estimation stage, which is applied to a pyramid of input images (video frames), as in [16]. Our method is then applied to the resulting video, where a pyramidal representation of the video frames is also used for the local motion estimation [17]. It should be noted that it is assumed that the video has been previously segmented into shots, so it is realistic to assume that the camera motion can be compensated for, since there will not be completely new frames during the sequence (as would be the case, e.g., after a hard cut in a video). Thus, after camera motion compensation, the (local) velocities in the video are estimated and are processed using higher order statistics, in order to determine which pixels undergo motion. This leads to the derivation of areas of activity, which are essentially the signatures of the activities taking place over the frames being examined. The activity areas are a good source of information for the type of events in the video, but they do not localize the moving entity with accuracy. They also do not contain any information about the appearance of the scene, which can be a rich source of semantics.

For this reason, we propose to take advantage of the color in the video sequence, and fuse it appropriately with the extracted activity areas, in order to use the data available in all domains and extract a richer set of semantics. Thus, at a second stage we perform color processing of the video, in order to describe its appearance and extract the moving entities. The background and the areas of activity are segmented based on their color, using mean shift segmentation. The resulting color layers in each area are compared, to find which pixels inside the activity areas belong to a moving object and which belong to the background. We present two different methods for comparing the color layers of the activity and background areas. The first method uses the Earth Mover's Distance (EMD) to compare the histograms of the color layers, and matches those (layers) that have the smallest distance between them. The second method is a novel approach, which is based on the statistical modeling of the colors in each layer. An appropriate probability density is fit to the colors of the activity and background areas, and a likelihood ratio test is formed to compare the respective probability densities. This creates a mask for each pixel, which contains its likelihood of belonging to the corresponding background layer. By thresholding this mask we can extract the moving object from the video frames.

The information extracted at this stage can be used itself to infer semantics concerning the video. For example, a tennis court can be identified based on color alone, or the place where the event is taking place can also be determined from its appearance, extracted during the color processing stage. The fusion of the motion with the color information leads to the segmentation of the moving objects, which gives the most complete description of their appearance, and can consequently result in the corresponding semantics, such as who is participating in the video. The likelihood maps show the likelihood ratio values comparing a pixel's color to that of the moving or static areas. This increases our system's flexibility, as we can choose to keep only pixels with high likelihood in the segmented objects, or also include the pixels with low probability.

This paper is organized as follows. The motion processing stage used to find the activity areas from the video is described in Section II. The color analysis method employed for the color segmentation of each frame is presented in Section III. The EMD and the LRT based methods that are employed for the comparison of the color layers are analyzed in Section IV. Experiments with real video sequences are presented in Section V. A quantitative comparison of the proposed methods with other combined color-motion approaches is provided in Section VI. Finally, conclusions and plans for future work are drawn in Section VII.

II. Motion Analysis: Activity Area extraction from optical flow

Optical flow techniques compute the illumination variations between pairs of frames, and attribute the changes in luminance to motion in the corresponding pixels [17], [18], under the constant illumination assumption. In this paper, the motion estimation is based on the estimation of the velocities between pairs of frames using a multi-resolution extension of local optical flow estimation methods, namely the Lucas-Kanade optical flow algorithm [19]. Local optical flow methods, which search for changes in illumination over local search windows, rather than entire video frames, are preferred in this work over global techniques, such as the Horn Schunck [20] method and extensions of it [21], [22], as they are more robust to noise [23], [24], and more efficient computationally. However, the success of Lucas-Kanade flow estimation is highly dependent on the size of the local search window used, and can fail in the case of large displacements. These limitations are overcome by resorting to a multi-scale implementation, which leads to improved accuracy in the flow fields. A coarse-to-fine approach can handle large displacements [25], because they are essentially translated to small displacements in the sub-sampled video frames. Indeed, usage of pyramids has been documented to provide much more accurate optical flow in areas of large displacements. Multi-resolution Lucas Kanade also achieves sub-pixel accuracy, through bilinear interpolation, thus all ranges of displacements are estimated reliably.

Although the resulting flow estimates are significantly improved, in comparison to nonpyramidal local flow estimation methods, they still provide more reliable results in textured areas of motion, whereas outliers appear at moving object boundaries, where the brightness constancy assumption is clearly violated. This drawback will be accounted for and overcome by our method, by accumulating flow estimates throughout video subsequences, to create “activity areas”. The outlier flow values at object boundaries are incorporated in the activity areas, so these areas also contain pixels that did not actually move, but were at object boundaries. However, these pixels will be removed along with all static pixels in each frame, in a color processing stage, which refines the results of the optical flow processing (Sec. III, V).

The constant illumination assumption of the optical flow approach is well known to be unrealistic. In real videos, there are slight illumination changes that may be introduced by camera instability, non-constant camera exposure, and other sources of measurement noise [24]. These illumination variations are often mistaken as motion, so the resulting optical flow estimates are noisy. When the flow is accumulated over a sequence of frames, some of the estimates are actual motion vectors, and some are measurement noise. The measurement noise affects all frame pixels, so a large number of samples of this random variable is available. As a consequence, we can satisfactorily model it by a Gaussian distribution [26], [27]. However, the inter-frame velocity estimates that are indeed caused by motion deviate from the Gaussian model, in most cases significantly, since the object motion is quite different from random illumination variations and measurement noise. The velocity estimates in frame *k* then correspond to the following two hypotheses:

$$\begin{aligned} H_0 &: v_k^0(\bar{r}) = z_k(\bar{r}) \\ H_1 &: v_k^1(\bar{r}) = u_k(\bar{r}) + z_k(\bar{r}). \end{aligned} \tag{1}$$

Hypothesis H_0 expresses a velocity estimate at pixel $\bar{r}$, in frame *k,* which is introduced by measurement noise, and hypothesis H_1 corresponds to the case where there is motion at pixel $\bar{r}$, expressed by the velocity $u_k(\bar{r})$, which is also corrupted by additive noise $z_k(\bar{r})$ [28]. In order to detect which velocity estimates correspond to a pixel that is actually moving, we need to examine the non-gaussianity of the accumulated velocity estimates [29]. A random variable’s non-gaussianity can be tested using higher order statistics, and, specifically, its kurtosis, defined as follows:

$$kurt(y) = E\{y^4\} - 3(E\{y^2\})^2 . \tag{2}$$

The fourth moment of a Gaussian random variable is $E\{y^4\} = 3(E\{y^2\})^2$, making its kurtosis equal to zero. The kurtosis is chosen to detect areas of motion in video sequences, as it has been shown to be a more robust detector [15] than simple differencing, which can suppress noise from Gaussian, and even non-Gaussian data, as shown analytically in [30]. Thus, it can be used in our application even for the cases where the noise in the optical flow estimates deviates from the Gaussian model assumption. Simple frame (or flow estimate) subtraction leads to noisier estimates for the areas of motion, as it cannot eliminate large pixel differences, which are due to heavy-tailed noise. The kurtosis can handle such problems, as it is a better measure of a distribution's "peakiness" [30]. This is verified by our experimental results as well, in Sec. V.

Thus, after collecting the noisy inter-frame velocity estimates of each pixel over several video frames, we estimate their kurtosis. We need to select a number of frames over which at least one event (motion) has occurred. For this reason, we initially collect a fixed number of frames, and with every new frame, we compare the new flow estimates with the mean of the previously collected ones. A "significant" amount of activity, i.e. a new event, has occurred when the new flow estimate is higher than the standard deviation of the previous ones. Then, we can either stop accumulating frames and extract the activity area, or we can gather more frames, to extract a region corresponding to additional motions. The pixels whose flow has kurtosis above 10% of the mean kurtosis are considered to have been displaced and form a mask for the region of each frame where motion has occurred. This leads to the formation of "activity areas", since we determine which pixels have moved in the frames that we are examining, but not the exact moving object. The activity areas can be useful for the extraction of semantic information concerning the sequence being examined, if, for example, they are characterized by a shape representative of specific actions.

A representative example of a boy playing tennis is shown in Fig. 1(a). Fig. 1(b) shows the activity areas extracted from processing frames 10 – 20. In this subsequence the boy has hit the ball, so his arm with the racket has moved, and the activity area contains the ball trajectory, along with the other areas of the boy's body that moved during these frames. The activity area of Fig. 1(c) has resulted from processing frames 1 – 50, where more parts of the player were moving. The trajectory of the ball arriving and leaving after being hit is clearly visible. Also, the shape of both activity areas shows the region where the tennis racket was moving as the boy hit the ball (the curve on the left of the boy's body before the ball was hit and on the right after the ball is hit).

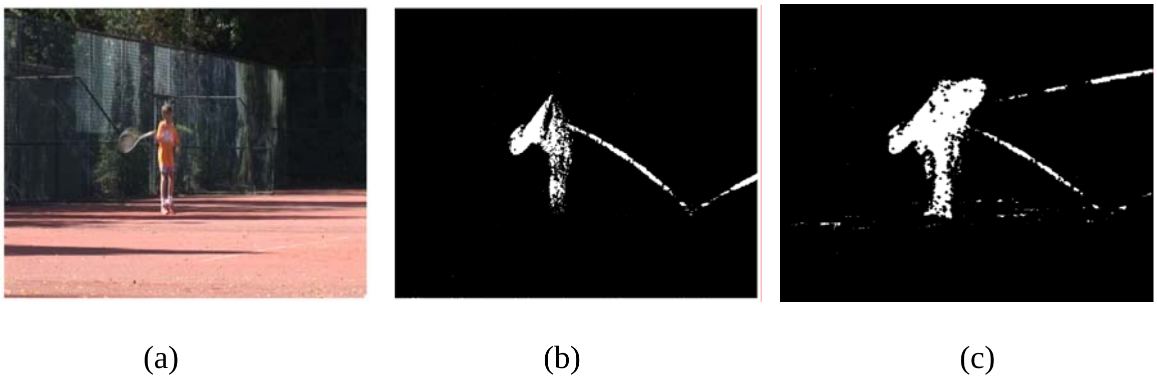

(a) (b) (c)

Fig. 1. Boy playing Tennis. (a) Frame 20. Activity areas: (b) frames 10 – 20, (c) frames 1 – 50.

III. COLOR SEGMENTATION: MEAN SHIFT

As seen in the previous section, the motion information can indicate what kind of events are taking place in a video. However, the extracted activity areas are not sufficient to fully characterize a scene, e.g. who is participating in an event, or where it is located. A much better understanding of the semantics in a video can be acquired by processing the color information available in it. The color alone may provide important information about the scene [31], for example a green area in a tennis game indicates a grass court, whereas a red one indicates a clay court [32]. By fusing the color with a scene's activity area, we can both segment the moving objects, and also extract additional semantic information concerning the video under examination. For example, an object's appearance is fully described after it is segmented, and the background area's color can indicate the place where the scene is being filmed. A significant advantage of the proposed method is that the color segmentation takes place on each frame separately, so the background modeling, for the extraction of the moving objects, is robust to variations in frame illumination.

We perform color segmentation on the video frame in order to separate it into regions of similar color. An elegant solution for the clustering of colors is provided by the mean shift algorithm [33]. The mean shift automatically defines the number of clusters, as it is an unsupervised learning algorithm, as opposed to the commonly used K-means [34]. Consequently, autonomous color clustering is a natural application for the mean shift [35]. The resulting clusters can be arbitrarily shaped, since no constraint is imposed on the cluster boundaries. The central idea of the mean shift algorithm is to shift a window of fixed size to the mean of the points in it. Mean shift essentially performs mode seeking, i.e. it searches for the maxima of the data distribution [36]. In our application, the data is appropriately modelled by a density function (Eq. (3)), and we search for this density's maxima (modes) iteratively, by following the direction where its gradient increases [33], [37]. This is achieved by iteratively estimating the data's mean shift vector (Eq. (5)), and translating the data window by that quantity, until it converges. It should be noted that convergence is guaranteed, as proven in [38].

For color segmentation, we convert the pixel color values to L*u*v* space, as distances in this space correspond better to the way humans perceive distances between colors [39], [40]. Thus, each pixel is mapped to a feature point, consisting of its L*u*v* color components, which we symbolize as **x**, in *d*-dimensional Euclidean space R^d, where d = 3 corresponds to the three color components. We consider that our data (consisting of $N_1 \times N_2$ 3D values **x**, for a $N_1 \times N_2$ video frame) can be modeled by a multivariate probability density function, which is expressed via a kernel *K*(**x**), and window of radius *h*, given by [36]:

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3}$$

where x_i represent the color values around which the distribution is centered. For the case of color segmentation $n \leq N_1 \times N_2$ points x_i are initially randomly selected from our data set. The kernel *K* of Eq. (3) is chosen to be symmetric (around x_i) and differentiable, so that it can enable the estimation of the pdf's gradient, and also its modes. The Epanechnikov kernel that we use is given by:

$$K_E(x) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1 - x^T x), & if \quad x^T x < 1 \\ 0, & otherwise \end{cases} \tag{4}$$

In [38] it is shown that, for the Epanechnikov kernel, we need to translate the window center x_i by the “sample mean shift” $M_h(x)$ at every iteration, in order to converge to the distribution’s modes. The sample mean shift is given by

$$M_h(x) = \frac{1}{n_x} \sum_{x_i \in S_h} x_i - x, \tag{5}$$

where n_x is the number of points contained in each search area $S_h(x)$. The mean shift $M_h(x)$ points in the direction of gradient increase, so it leads to the pdf maxima (modes) of our data.

Thus, in order to obtain the color segmentation of each video frame, we first convert the image into L*u*v* space. We randomly choose *n* image feature points x_i in this space as initial cluster centers, and for each $i = 1, \dots, n$, we estimate the sample mean shift $M_h(x_i)$ in a window $S_h(x_i)$ of radius *h* around point x_i. The window $S_h(x_i)$ is translated by $M_h(x_i)$, and a new sample mean shift is estimated, until convergence, i.e. until the shift vector becomes approximately zero. Finally, the pixels with color values closest to the density maxima derived by the mean shift iterations are assigned to those cluster centers.

IV. Fusion of Activity Areas and Color for Moving Object Segmentation

By applying the mean shift algorithm, each frame is separated into color-homogeneous regions, as in [38], [39]. We use this result to segment the objects of similar color in the video frames, regardless of their shape. The motion information processing has led to the separation of active pixels in the video from the pixels that do not undergo any motion. By applying the mean-shift based color segmentation in the areas where no activity occurs, we can determine which colors are present in the background. Similarly, the color segmentation of the activity areas allows us to separate them in color homogeneous regions, that correspond to both the moving object and the background. Consequently, by matching each color homogeneous regions of the background to the corresponding regions in each frame’s activity area, we can determine which pixels of the activity area are background.

In Fig. 2 we show the mean shift color segmentation of frame 36 (Fig, 1(a)), where it is clear that the boy’s colors have been separated from the colors of the tennis court. Fig. 2(b) shows the color segmentation of the activity area, where, again, the boy’s colors are separated from the background. Finally, the color segmentation of the static regions is shown in Fig. 2(c), where the different colors in the background have been separated. The extracted colors of each area need to be matched in a reliable manner, in order to accurately retrieve the pixels corresponding to the player’s location in each frame.

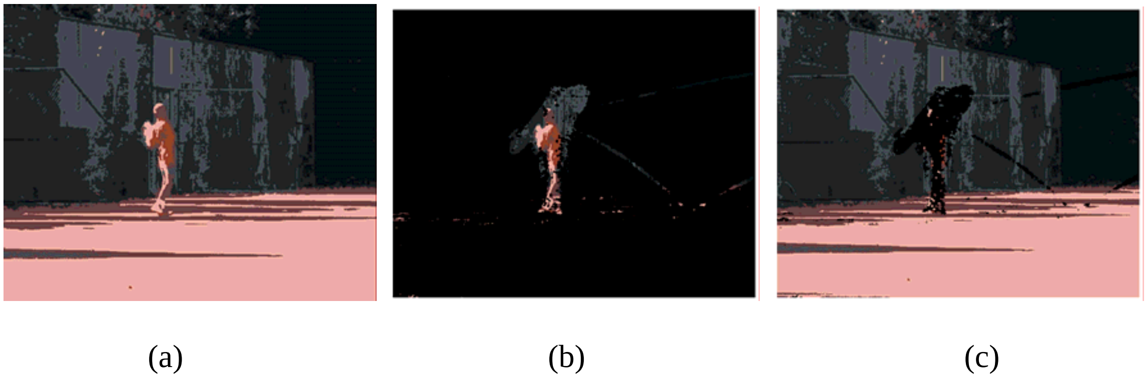


Fig. 2. Mean shift color segmentation: (a) Frame 36, (b) activity area, (c) background.

A. *Earth Mover's Distance*

Many methods exist for the comparison of areas of different colors, in our case of the color regions in the activity areas and the background. Each color region is represented by three histograms corresponding to its three color components. The histograms of each color can be regarded as "signatures" characterizing its distribution. The similarity between signatures of data is measured by the Earth Mover's Distance (EMD) [41], that calculates the cost of transforming one signature to another. The most similar signatures will have the lowest cost and, consequently, can be clustered together. In this paper, histograms will be used as signatures, as the large amount of data processed in a video makes the use of more sophisticated signatures (e.g. histograms with adaptive binning) computationally impractical.

A histogram with *m* bins can be represented by $P = \{(\mu_1, \Sigma_1, h_1), \ldots, (\mu_m, \Sigma_m, h_m)\}$, where μ_i, Σ_i are the mean and covariance, respectively, of the data in that bin (equivalently, cluster), and h_i is the corresponding histogram value (essentially the probability of the values of the pixels in that cluster). This histogram can be compared with another, $Q = \{(\mu_1, \Sigma_1, h_1), \ldots, (\mu_m, \Sigma_m, h_m)\}$, by estimating the cost of transforming histogram *P* to *Q*. If the distance between their clusters is Δ_{ij} (we use the Euclidean distance here), the goal of transforming one histogram to the other is that of finding the flow f_{ij} that achieves this, while at the same time minimizing the cost:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} \Delta_{ij} f_{ij}. \tag{6}$$

After the optimal flow f_{ij} between clusters *i* and *j* is found [41], the EMD becomes:

$$EMD(P;Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \Delta_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}. \tag{7}$$

In order to compare color distributions, the multi-dimensional color histogram can be used. However, estimating the precise, joint color distribution of each color cluster is computationally demanding. The subsequent comparison of these multi-dimensional distributions further increases the computational cost. For the case of video processing, in particular, this can become computationally prohibitive, as the colors histograms of all segmented areas, in all video frames need to be compared. We are thus led to treat the three color components of each image area as uncorrelated and independently distributed. This assumption is made in the literature for kernel color density estimation [42], [43], and for computationally efficient color modeling by mixtures of Gaussians [44]. In practice, both in the literature and in our experimental results, this suboptimal independence assumption leads to good results. Nevertheless, examining the use of more precise color models, that are also computationally efficient, is also possible.

Thus, we estimate the EMD between the three histograms of each color region in the action mask and the background area of each frame. We combine the EMD's results for each color component (L, u, v in this case) by simply adding their magnitudes. The pixels of each color region in the activity area and the background that require the least cost (EMD) to be transformed to each other correspond to the background pixels of the activity areas. By removing the "background pixels" from the activity areas, we aim to extract the moving objects from the video.

B. Probability Likelihood Testing for Color Comparison

The EMD can become computationally expensive, especially for large amounts of data, like those encountered in video applications. This motivates us to develop an alternative approach for the comparison of the color regions. Our method leads to the extraction of the moving objects by comparing the color in the activity area with the background color, but it also has the additional advantage of providing an estimate of each pixel's likelihood to belong to the moving object, instead of only providing a binary decision (like EMD-based comparison does). We estimate the statistical distribution of each color homogeneous layer in the activity and the background areas, and compare these two distributions. This is equivalent to forming a hypothesis test for each pixel $\bar{r}$, with color $\bar{x}$ (there are three color components, e.g. R, G and B or L, u and v, so $\bar{x}$ is 3D), with the hypotheses that it has moved in the current frame or is a static pixel:

$$\begin{aligned} H_0 &: \bar{x} \sim f_{static}(\bar{x}) \\ H_1 &: \bar{x} \sim f_{active}(\bar{x}). \end{aligned} \tag{8}$$

This process requires modeling of the color data's statistical distribution for the pixels inside the activity area (f_{active}), and its complement (f_{static}). As in Sec. IV-A, we make the assumption that the three color components in each region are independent random variables [44], [43], so the overall pdf of pixel $\bar{r}$ can be written as:

$$f_{area}(\bar{x}) = f_{area,L}(\bar{x})\, f_{area,u}(\bar{x})\, f_{area,v}(\bar{x}), \tag{9}$$

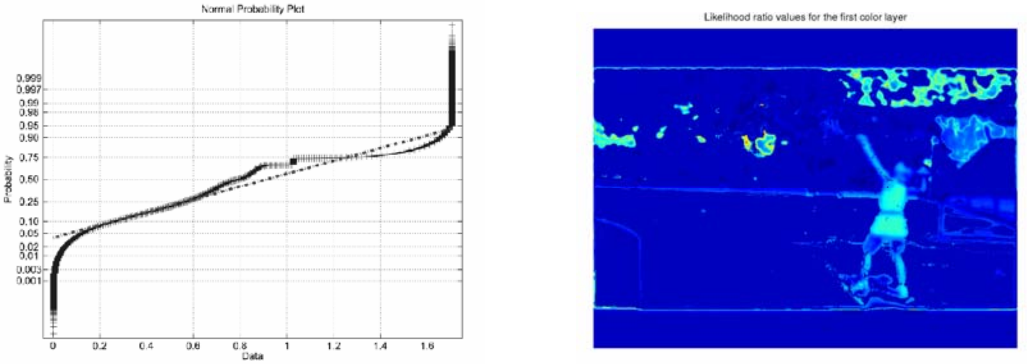
where "area" is static or active. Spatial luminance data is often modeled by mixtures of Gaussian distributions, which are estimated using the Expectation-Maximization (EM) algorithm [45]. However, the EM requires training, which in turn would require either prior knowledge of which frames (and which regions of the frames) are appropriate for training, as well as more computational overhead. The EM also needs knowledge of the number of mixture components, which is usually determined in an ad-hoc manner. Its success is highly dependent on the number of mixture components and, most of all, on correct training, which limits its usability in general applications. For these reasons, we develop a simpler but effective method for approximating the probability distribution of each color layer, and subsequently comparing the different color regions.

The normal probability plot (NPP) plots the "percentage points" of the normal distribution, which are linearly related to the ordered data values (when these follow a normal distribution). The "percentage point" for each ordered value j is given by $z_j = G(p_j)$, where G is the inverse cumulative distribution function, and $p_j = (j - 3/8)/(N + 1/4)$ for N samples [46]. The horizontal axis shows the ordered sample data values. Thus, the NPP of normally distributed data should follow a straight line when the ordered values j are plotted against the percentage points, and deviations of the plot from a straight line indicate deviations of the data from a Normal distribution.

The empirical cumulative probabilities for the data under examination are estimated and superposed on the NPP to see if they indeed follow a Normal distribution. We normalize our data by subtracting its mean and dividing by its standard deviation, so if it follows a Gaussian distribution, its normalized values should follow a Normal distribution. The NPP of Fig. 3(a) shows that our (normalized) data has empirical probability values that deviate from the straight line, and consequently cannot be accurately modeled by a Normal (or Gaussian) distribution. This is because the color values contain outliers that introduce heavier tails than the Gaussian distribution. We account for the data's outliers by using the heavy-tailed Cauchy distribution, given by:

$$f(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - \delta)^2}, \tag{10}$$

where γ is the data's dispersion and δ is its location parameter (the median for the Cauchy pdf). The dispersion parameter of the Cauchy distribution expresses the data's spread around its median δ, so its meaning is similar to that of the data variance for distributions where it is defined (the variance is not defined for the family of stable distributions, to which the Cauchy distribution belongs) [47].



(a) (b)

Fig. 3. (a) Normal Probability Plot for a frame's color layer. (b) Likelihood ratio values.

A likelihood ratio test (LRT) is then formulated, to find whether each frame's pixel belongs to the color layer of the action or the static area. Using the Cauchy model for our data and Eq. (9), we have:

$$\Lambda(\bar{x}) = \frac{f_{active}(\bar{x})}{f_{static}(\bar{x})} = \prod_{i=\{L,u,v\}} \frac{\gamma_{static,i}}{\gamma_{active,i}} \frac{\gamma_{active,i}{}^{2} + (x - \delta_{activec,i})^{2}}{\gamma_{static,i}{}^{2} + (x - \delta_{static,i})^{2}}, \tag{11}$$

where $\bar{x} = \lfloor x_{L,} x_{u,} x_{v} \rfloor$ are the examined pixel's luminance values (in this case in L*u*v* space), layer = {static or active}, and the parameters $\gamma_{layer,i}$ and $\delta_{layer,i}$ are estimated directly from the data available [47]. In order to mask out the player, we threshold the values of the LRT, using the Bayesian threshold [28], which is automatically extracted from the data:

$$\eta = \frac{\mu_{H_1} + \mu_{H_0}}{2}, \tag{12}$$

where μ_{H_1} and μ_{H_0} are the means of the LRT of Eq. (11). They are directly estimated from the data available as follows: $\mu_{H_1} = E_{H_1}[L(\bar{r})]$, $\mu_{H_0} = E_{H_0}[L(\bar{r})]$, where, for H_1 we estimate the LRT mean using pixels in an active area, and for H_0 we use the rest of the frame pixels. By thresholding the LRT, we separate the moving object pixels from the background pixels in each frame. As our experiments show, this method gives equally good results as the EMD-based one, and in some cases performs even better, at a much lower computational cost and with a simpler implementation. Another advantage of using the LRT is that it gives a measure of the likelihood with which the pixels in the action mask match the corresponding color layer of the static regions, as shown in Fig. 3(b). This gives us the flexibility to decide whether the segmentation should include all possible object pixels, at the cost of including "false alarm" pixels, i.e. pixels that did not move at that frame, or including only the high LRT pixels, at the cost of losing some of the object pixels.

V. Experiments

In the following experiments, we demonstrate the results for various stages of our algorithm on real video sequences, where color and motion are used to achieve accurate segmentation and semantically meaningful results.

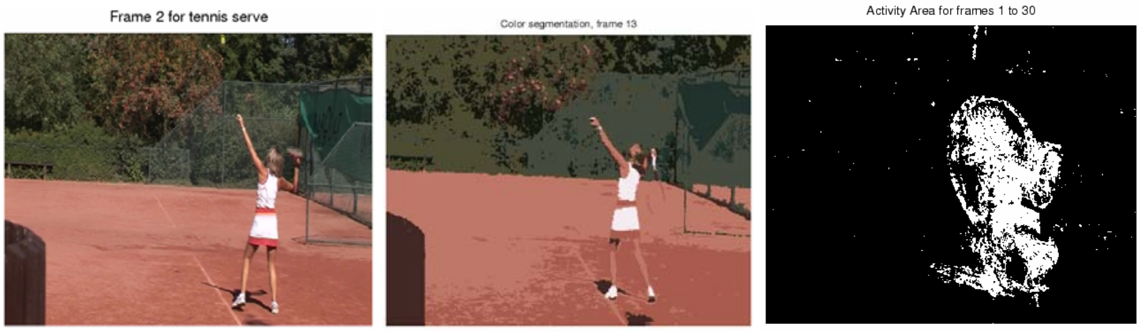
We also compare the results of our approach to other methods combining motion and color for the purposes of object segmentation. The estimation of activity areas using higher order statistics (the kurtosis) is compared to the activity areas extracted via simple frame differencing, as in the work focusing on motion energy images and motion history [48], [49]. Indeed, our method gives more reliable results for the regions of activity (Sec. II), as both quantitative and qualitative results show.

The proposed method is also compared against the methods in [15], [13]. In [15], mean shift color segmentation is first applied to the video frames, and motion is detected in color homogeneous regions, by estimating the fourth order moment of the illumination difference between regions of similar color. This approach does not estimate the optical flow, but simple frame-region differences, in order to reduce its computational cost. We expect that it will lead to less accurate activity areas than our flow based approach, since simple frame differencing is more susceptible to false alarms. Afterwards, an affine model is fitted to the motion of each color region, and similarly moving regions are clustered [15]. This limits the applicability of the method to color homogeneous moving objects, whereas our approach does not impose such restrictions on the objects to be segmented, and can therefore deal with a larger range of motions and object appearances. In [13], clusters with similar color and spatial location are grouped over a sequence of frames, for tracking purposes. However, this requires prior knowledge of the number of clusters, making it impractical in realistic applications. Additionally, by clustering pixels that have similar color and location, it cannot deal with moving objects that have very different colors, or have similar colors which appear in different locations (e.g. after sudden large displacements, which are common in sports). These motion and color based methods also have a higher computational burden than our method, as they require the estimation of displacements between regions of similar color, which is equivalent to estimating the motion between many sequences, formed by the color clusters from all video frames.

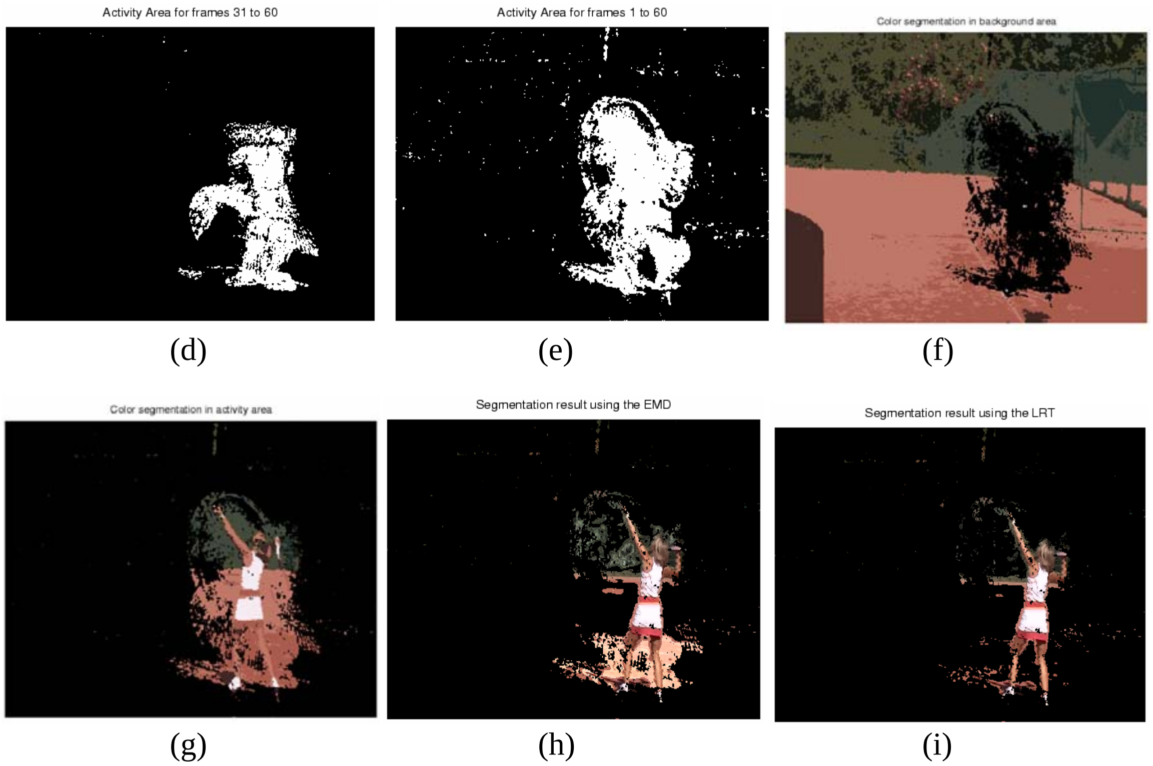
All these methods are applied on video sequences from various domains (sports, surveillance, nature etc), with varying degrees of complexity of color information and motion, and with multiple moving entities. Quantitative evaluation and comparison of the results is provided in Sec. VI and the processing times are compared. In the first experiment we present the analytical explanations and qualitative results of all methods, but to avoid repetition, we provide a table with quantitative comparison results of the quality of segmentation for all videos (including the first one).

A. Tennis serve

In this experiment, our method is applied to a color video of a tennis serve. In Fig. 4(a) the player is shown before she hits the ball, and Fig. 4(b) shows the results of the color segmentation. The optical flow is then estimated and processed as described in Sec. II, leading to the activity areas shown in Fig. 4(c) – (e). We show the results of accumulating the flow over different subsequences of video frames, to demonstrate that the resulting areas are representative of the actions taking place.


Frame 2 for tennis serve
Color segmentation, frame 13
Activity Area for frames 1 to 30

(a) (b) (c)


Activity Area for frames 31 to 60
Activity Area for frames 1 to 60
Color segmentation in background area
Color segmentation in activity area
Segmentation result using the EMD
Segmentation result using the LRT

(d) (e) (f)

(g) (h) (i)

Fig. 4.Tennis serve: (a) frame 2. (b) Mean Shift color segmentation for frame 13. Activity Areas for a Tennis Serve: (c) frames 1 – 30, (d) frames 31 – 60, (e) frames 1 – 60. Color segmentation for: (f) background, (g) activity area. Segmentation results: (h) EMD, (i) LRT.

By performing color segmentation in the resulting activity and background areas, we can acquire even more information about the video. The color alone can provide information about the kind of tennis court, whereas the matching of the color in the activity area and the background area will lead to the segmentation of the tennis player. Fig. 4(f), (g) shows the results of the mean shift color segmentation applied to the background, as well as the activity area in each frame.

By comparing the color layers using the EMD and the LRT, we obtain the segmentation results shown in Fig. 4(h), (i). The LRT gives better segmentation results than the EMD, and also provides us with the likelihood mask of Fig. 3(b). The better performance of the LRT-based method can be attributed to the fact that the test based on the EMD sums the absolute values of the EMD's between the three color components, whereas the LRT approximates the color distribution, with a model that is theoretically suboptimal, albeit sufficient for practical purposes.

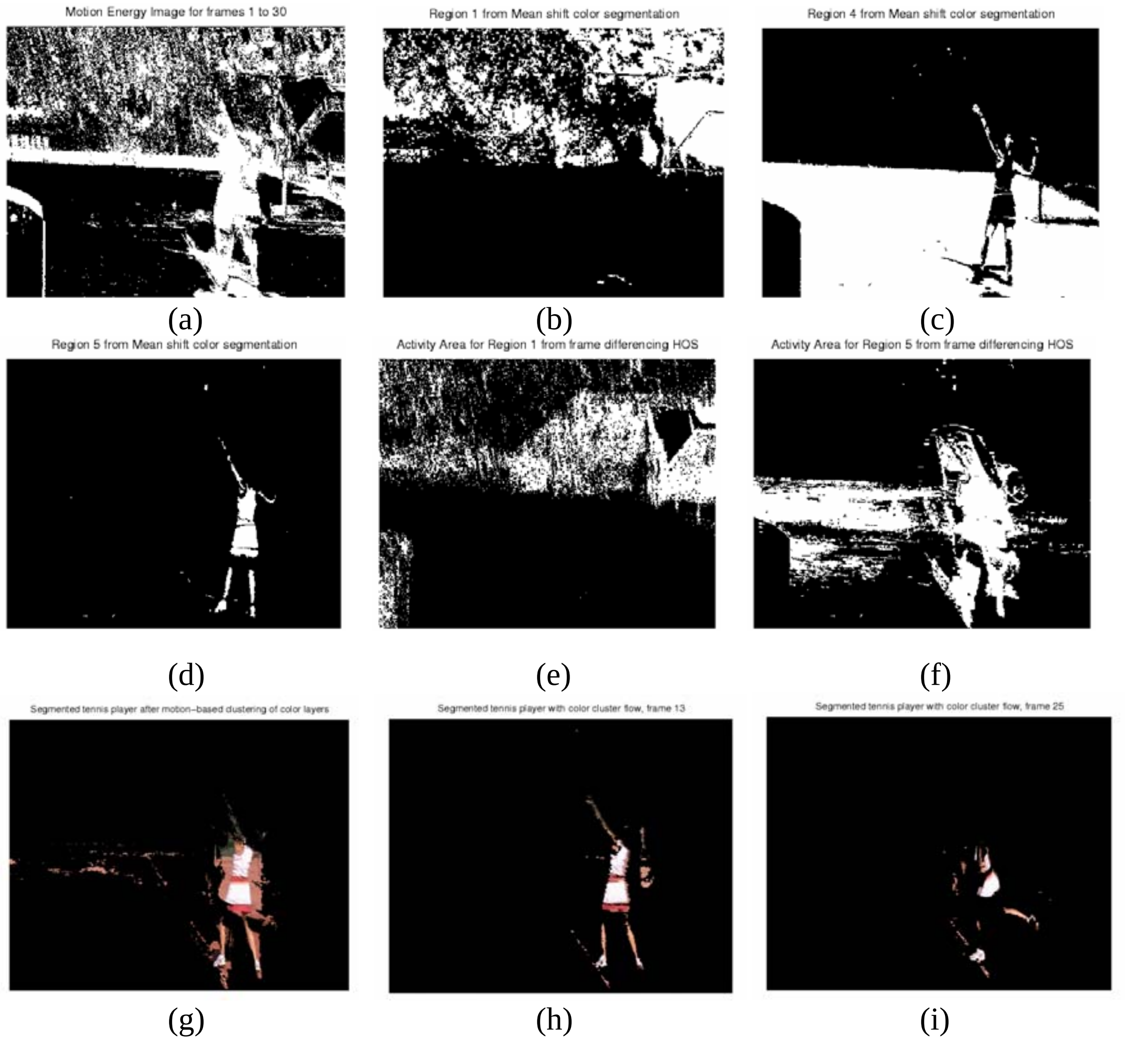
Analytical Comparison with color-motion based methods.

For this video sequence, we present an analytical comparison of its performance with other approaches, including qualitative results. Due to space limitations, and since the same analysis applies to the results from the other video sequences, a perceptually meaningful quantitative performance comparison of the segmentation results for this and the other videos is provided in Sec. VI.

In order to compare the performance of our approach to methods that extract foreground areas using inter-frame thresholded differences, we first extract the motion energy image, as in [49], by taking the union of the binarized (thresholded) frame differences. From Fig. 5(a) it is obvious that the resulting area of motion contains many errors, as it extracts an entire region of the background vegetation, caused by small leaf motions and camera instability. Our method avoids these errors (Fig. 4(e)), as it is based on higher order statistics of the optical flow, whose noise-induced variations in those locations are effectively eliminated. Thus, in the sequel simple frame differencing, as in [49], [48], is not examined, as it leads to less accurate results than the higher order statistics processing of the optical flow estimates.

In order to compare our approach with the work of [13], [15], we also examine the extraction of color homogeneous regions, and the subsequent estimation of their motion. In Fig. 5 (b) – (d), three examples of binary masks, corresponding to layers with similar color are shown (there are in total five such regions), corresponding to the color segmentation shown in Fig. 4(b). For comparison purposes, we extract areas that are equivalent to our activity areas, using the approach of [15]. In [15], the higher order statistics of the differences between color homogeneous layers (regions) of different frames lead to moving regions. A region-based affine motion model is then used to find the motion parameters for these regions, and cluster them. After morphological post-processing, this produces activity areas, as in Fig. 5(e), (f). The area in Fig. 5(e) is less noisy than the motion energy image [49], as it makes use of higher order statistics, applied to a specific region of the video frames. Nevertheless, there are still many errors, caused by error propagation from the initial segmentation stage, and false alarms introduced by small luminance changes in background areas, which are mistaken for motion. These false alarms are avoided in our approach, since we process the optical flow values, which are more robust to small illumination variations than simple frame differencing. In order to achieve object segmentation, the estimated affine models are used to track activity areas in the next frames [15]. When the color segmented area and the tracked area pixels overlap by more than 85%, they are considered to belong to the same object. The results of applying this approach, displayed in Fig. 5(g), are qualitatively worse than the proposed LRT-based approach (Fig. 4(i)). As the clustering groups the pixels in color layers which undergo similar motions, some parts of the player with little or different motion, such as the arms or the head, are not extracted. Both the EMD and LRT-based methods avoid this, since the activity mask they use is created by accumulating all the motions that take place over many frames.

In the Color Cluster Flow method of [14], the location of color segments is included in a feature vector, which is used to cluster similarly colored and located pixels in consecutive frames. An obvious initial drawback of this method is that it requires manual selection of the clusters to be tracked, when dealing with video. As Fig. 5(h) shows, this method gives good results when the moving objects have similar colors in nearby locations. Even so, the head of the tennis player is not included in the segmentation result, as it corresponds to a different color region. As expected, this method is unreliable for the case of non-rigidly moving objects, when they undergo large displacements. Fig. 5(i) shows a representative example of the same method, applied to frame 25 of the tennis sequence, where the player's legs and arms have moved significantly, and parts of them (and also the head) have been lost, as those pixels do not simultaneously satisfy the color and location similarity criteria. This, and the experiments with other sequences that follow, show that this method is not reliable for more challenging types of motion, or for moving objects with very different colors.


Motion Energy Image for frames 1 to 30

Region 1 from Mean shift color segmentation

Region 4 from Mean shift color segmentation

(a) (b) (c)

Region 5 from Mean shift color segmentation

Activity Area for Region 1 from frame differencing HOS

Activity Area for Region 5 from frame differencing HOS

(d) (e) (f)

Segmented tennis player after motion-based clustering of color layers

Segmented tennis player with color cluster flow, frame 13

Segmented tennis player with color cluster flow, frame 25

(g) (h) (i)

Fig. 5. (a) Motion energy image from union of thresholded inter-frame differences, from frames 1 to 30. Regions from mean shift color segmentation: (b) region 1, (c) region 4, (d) region 5. Activity area from higher order statistics of inter-frame differences: (e) in region 1, (f) in region 5. Segmentation results: (g) motion-based clustering of color layers, frame 13. Color-cluster flow: (h) frame 13, (i) frame 25.

B. Tennis hit

In this experiment, we examine a video of a tennis player hitting the ball (Fig. 6(a)), whose color segmentation is shown in Fig. 6(b). The activity areas extracted after processing the optical flow estimates for frames 1 – 10, 1 – 30 respectively, are shown in Fig. 6(c), (d). As expected, these signatures indicate the activity taking place in each subsequence, for example the trajectory of ball is visible, and the motions of the player's arms can also be seen.

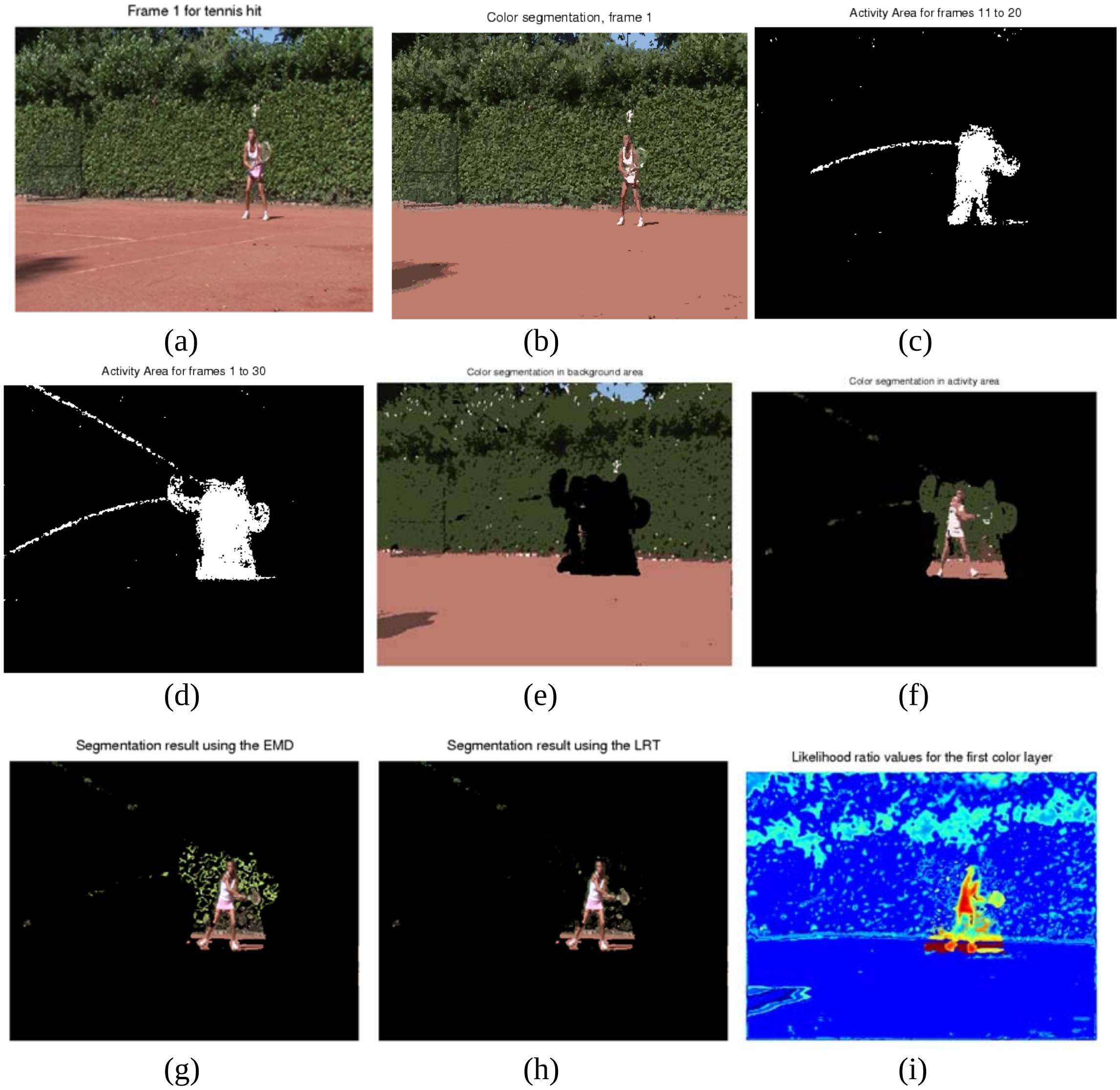
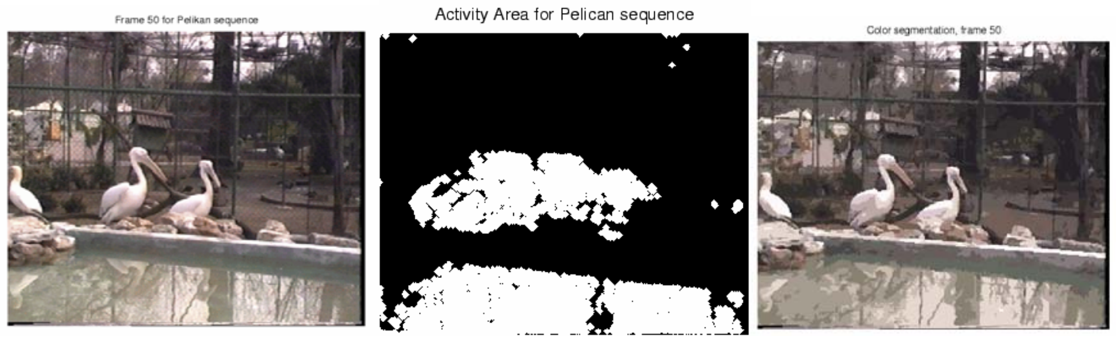

Frame 1 for tennis hit
Color segmentation, frame 1
Activity Area for frames 11 to 20
(a)
(b)
(c)
Activity Area for frames 1 to 30
Color segmentation in background area
Color segmentation in activity area
(d)
(e)
(f)
Segmentation result using the EMD
Segmentation result using the LRT
Likelihood ratio values for the first color layer
(g)
(h)
(i)

Fig. 6. Tennis hit: (a) frame 1. (b) Color segmentation. Activity Areas for Tennis Hit: (c) frames 1-10, (d) frames 1-30. Mean Shift Color Segmentation: (e) background, (f) activity area. Segmentation results: (g) EMD, (h) LRT. (i) Likelihood ratio values.

As before, the motion processing results are complemented by the mean shift color segmentation, so errors introduced in the activity areas, e.g. by outlier flow values, can be eliminated (Sec. II). In Fig. 6(e), (f) we see the results of color segmentation on the background and the activity areas of a frame of the video sequence. After comparing and matching the color histograms of the color layers in the segmented activity area and background area using the EMD and the LRT, we obtain the correct segmentation of the player, shown in Fig. 6(g), (h). As before, the result of the EMD method does not remove all the background pixels from the activity area. Also, the LRT method provides us with the values of the likelihood ratio for comparing the colors of the foreground and the background, shown in Fig. 6(i), where it can be seen that the player's colors are quite different from those of the background, and validate the LRT-based method's good performance.

C. Low Color Contrast Video: Pelikans

In this experiment we have applied the proposed method to a video of pelikans walking, where the colors of the background and of the pelikans are similar (Fig. 7(a)). The extracted optical flow is processed to obtain the activity area shown in Fig. 7(b). In this case the activity area includes the moving birds as well as the lake, because its water is rippling, and also because the moving pelikans are reflected in it. The color segmentation obtained using mean shift applied to the entire frame, the activity area and background region is shown in Fig. 7(c) – (e). After comparing the color layers inside the activity and background areas using the EMD and the LRT, we obtain the results of Fig. 7(f) – (i), where, again, the LRT-based method gave a more precise segmentation of the moving pelikans, whereas the EMD-based approach also included a part of the background behind them.


Frame 50 for Pelikan sequence
Activity Area for Pelican sequence
Color segmentation, frame 50

(a) (b) (c)

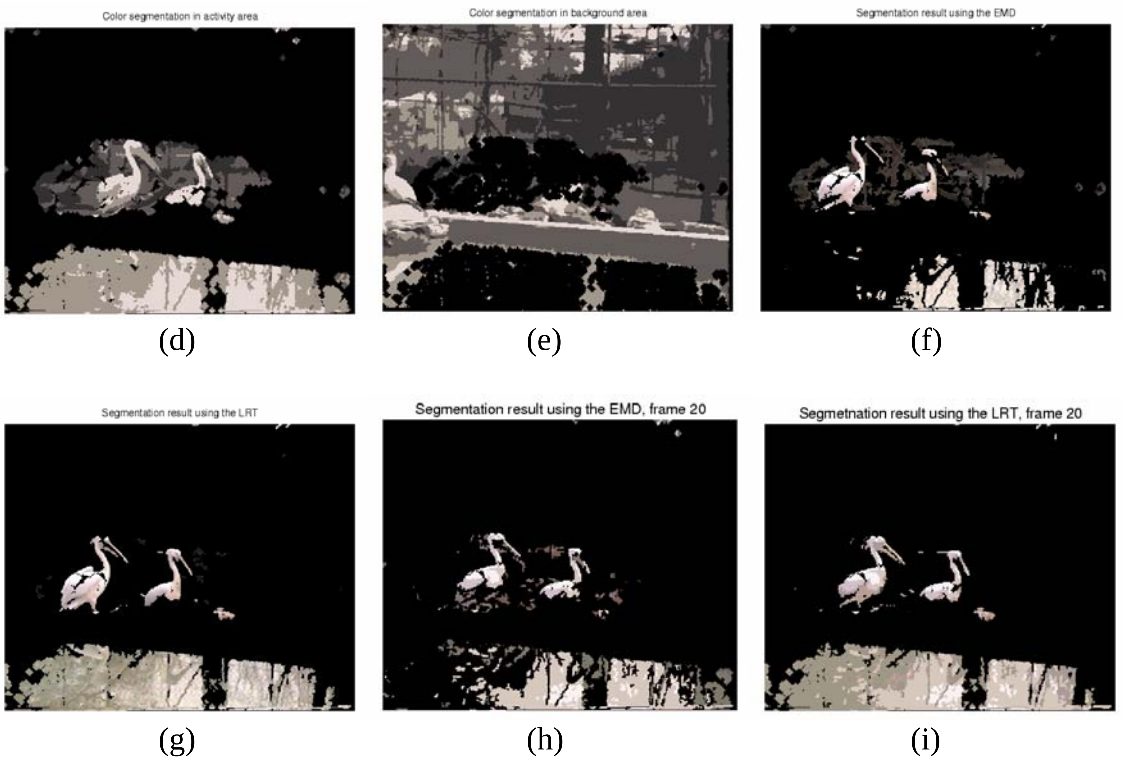
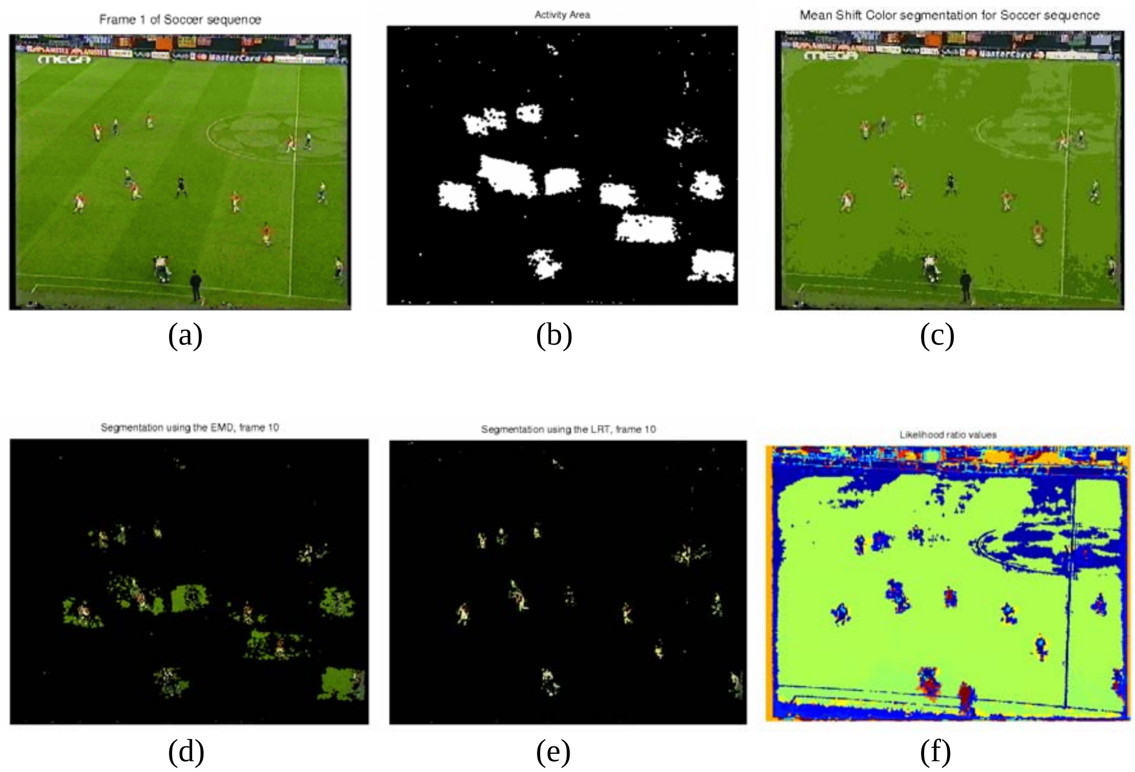

Color segmentation in activity area
Color segmentation in background area
Segmentation result using the EMD
(d)
(e)
(f)
Segmentation result using the LRT
Segmentation result using the EMD, frame 20
Segmetnation result using the LRT, frame 20
(g)
(h)
(i)

Fig. 7. Pelikan Sequence. (a) Frame 50. (b) Activity Area. Color segmentation for: (c) Frame 50. (d) Activity Area. (e) Background Area. Segmentation results. Frame 7: (f) EMD, (g) LRT. Frame 20: (h) EMD, (i) LRT.

D. Many Moving Objects: Soccer sequence

In order to examine the robustness of our method when the number of moving objects increases, we have applied it to a soccer sequence, where many players are moving in various directions (Fig. 8(a)). The activity area extracted for this sequence is shown in Fig. 8(b), where it is already obvious that the areas where the players are moving have been successfully localized. The mean shift color segmentation is shown in Fig. 8(c), where the colors of the players and the field have been effectively separated. The color layers in the activity and background areas are compared using the EMD and LRT-based methods in order to separate the players from the soccer field. The results are, as in the previous cases, better for the LRT-based method, which extracts the players with more precision, as shown in Fig. 8(d), (e). The likelihood ratio values in Fig. 8(f) are clearly higher in the players' pixels than the background pixels, as expected. It should be noted that the proposed method extracts the moving entities in each video. In Fig. 8(c) it is evident that the immobile observer (the dark silhouette in the bottom of this image) has been successfully segmented from the background, based on color, but the activity area of Fig. 8(b) eliminates him from the final segmentation result (Fig. 8(d), (e)), as he is not moving. Further processing could also enable the extraction of all people on the soccer field, even if they are motionless. Specifically, the comparison of background pixels with the segmented moving objects, based on color, would enable the extraction of this person as well.


Frame 1 of Soccer sequence
Activity Area
Mean Shift Color segmentation for Soccer sequence
Segmentation using the EMD, frame 10
Segmentation using the LRT, frame 10
Likelihood ratio values

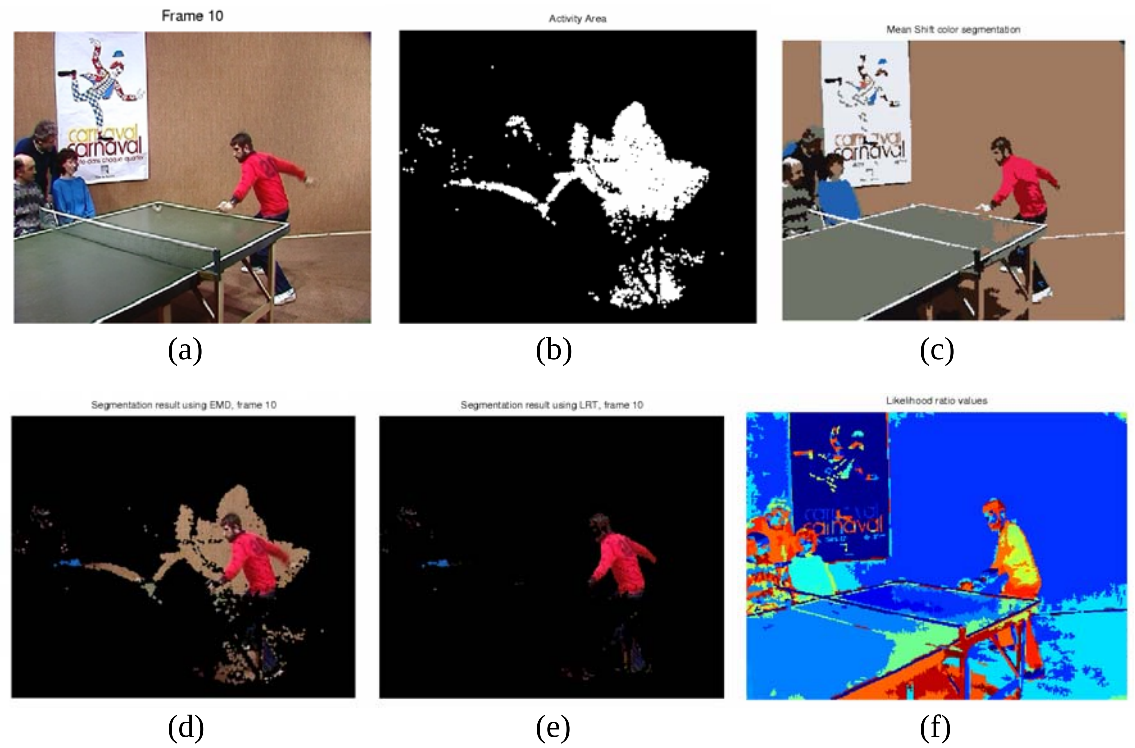
(a) (b) (c)

(d) (e) (f)

Fig. 8. Soccer Sequence. (a) Frame 1. (b) Activity Area. (c) Mean Shift color segmentation. Soccer Sequence segmentation results. (d) EMD-based for frame 10. (e) LRT-based for frame 10. (f) Likelihood ratio values.

E. Table Tennis

Experiments have been conducted with a video of a person playing table tennis (Fig. 9(a)). The player's motion is highly non-rigid, as the resulting activity mask shows in Fig. 9(b). The mean shift color segmentation in Fig, 9(c) shows the color layers, where some of the colors in the foreground area are similar to those in the background. The proposed algorithm, using the EMD color-comparison does not give an accurate segmentation of the player, as shown in Fig. 9(d), whereas the LRT-based method leads to a much more precise segmentation, in Fig. 9(e). It should be noted that in the results of the LRT-based approach, the player's face is not extracted, as its color is very similar to that of the background. Finally, the likelihood ratio values are displayed in Fig. 9(f), where it can be seen that the colors of the player lead to higher likelihood ratio values in the correct pixels.


Frame 10
Activity Area
Mean Shift color segmentation
Segmentation result using EMD, frame 10
Segmentation result using LRT, frame 10
Likelihood ratio values

(a) (b) (c)

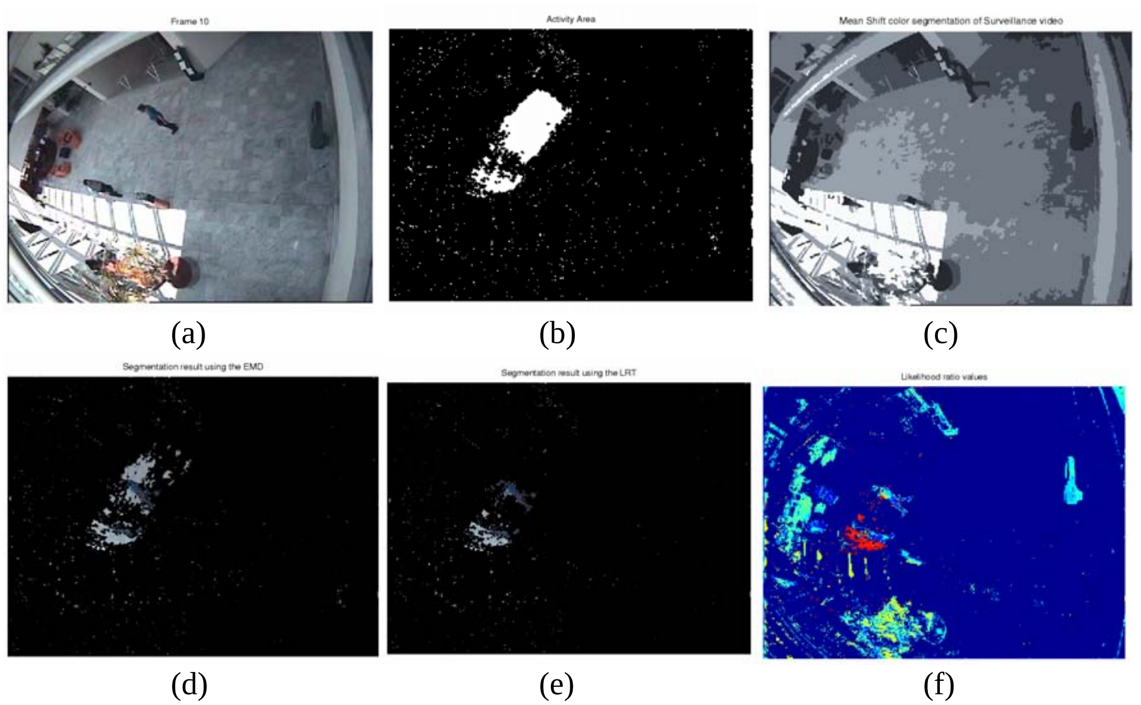
(d) (e) (f)

Fig. 9. Table Tennis. (a) Frame 10. (b) Activity Area. (c) Mean Shift color segmentation. Segmentation results. (d) EMD-based for frame 10. (e) LRT-based for frame 10. (f) Likelihood ratio values.

F. Surveillance

We also perform experiments with a surveillance video, taken from the PETS-ECCV’04 benchmark data (available online at http://www-prima.imag.fr/PETS04/caviar data.html), which shows two people meeting inside a building. Fig. 10(a) shows a frame of the video and Fig. 10(b) shows the resulting activity area. The large white area on the right shows that that person is walking towards the other for a longer time. Information like this can prove very useful in applications that are trying to extract semantics, such as detecting suspicious activities from surveillance videos.

The mean shift color segmentation is shown in Fig. 10(c), where it can be seen that parts of the people walking have a different color from their environment and can be successfully separated. Other regions (like their shirts) are similarly colored to their environment, and could be confused with it. This difficulty is overcome by taking advantage of the motion information, which limits the search area, and consequently the chances of false alarms (Fig. 10(b)).


Frame 10
Activity Area
Mean Shift color segmentation of Surveillance video
Segmentation result using the EMD
Segmentation result using the LRT
Likelihood ratio values

(a) (b) (c)

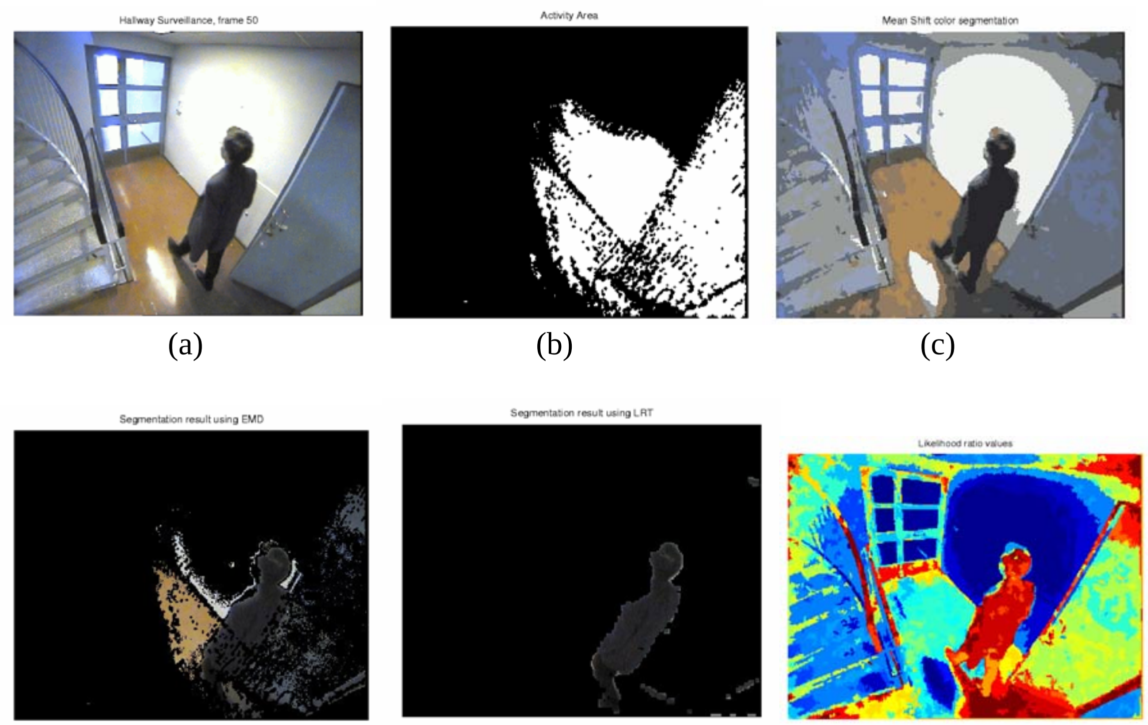
(d) (e) (f)

Fig. 10. Surveillance. (a) Frame 10. (b) Activity Area. (c) Mean Shift color segmentation. Segmentation results: (d) EMD, (e) LRT. (f) Likelihood ratio values.

Despite this, the person that is walking from the left is not segmented out as accurately as the person on the right. The color histograms of the resulting color layers are then matched using the EMD and the LRT methods, and the results are shown in Figs. 10(d) and (e). As in previous cases, the LRT-based method has produced better segmentation results, as the EMD results contain more regions of the background. Finally, the likelihood of a pixel to belong to the moving objects is shown in Fig. 10(f). The likelihood is higher in the regions corresponding to the two people, but also contains false alarms in the regions under the windows. However, the incorporation of the motion information via the activity mask eliminates these areas, as no motion was detected in them. Thus, we see that the complementary roles of motion and color reduce the errors that each method alone might introduce, and leads to good segmentation results. The color and motion based techniques of [13] and [15] give worse results for this video, as the initial color segmentation is not sufficient for accurate segmentation and tracking of the two people (Sec. VI).

G. Surveillance - Hallway

We applied the proposed hybrid color-motion approach to a surveillance video of a person entering a hallway (Fig. 11(a)). In this case the motion of the person is non-rigid, and he undergoes perspective transformation, as he walks away from the camera. Also, the shape of the person changes, as different parts of him become visible, and his color is similar to some parts of the background (especially the staircase railing and the shade). The activity mask for this video is shown in Fig. 11(b), and includes the parts of the scene where the door is opening, and the person is walking. The color segmentation of Fig. 11(c) clearly separates the different color regions. By comparing the color in the activity and background areas, we obtain the segmentation results of Fig. 11(d), (e), where, as before, the EMD-based results are generally noisier than the LRT-based ones, as they include more areas of the background. The quantitative comparison of the proposed system with other color-motion methods in Table I also shows that it outperforms them.


Hallway Surveillance, frame 50
Activity Area
Mean Shift color segmentation
Segmentation result using EMD
Segmentation result using LRT
Likelihood ratio values

(a) (b) (c)

(d) (e) (f)

Fig. 11. Surveillance of hallway. (a) Frame 50. (b) Activity Area. (d) Mean Shift color segmentation. Segmentation results: (e) EMD, (f) LRT. g) Likelihood ratio values.

VI. Quantitative Comparison

In order to evaluate the quality of the segmentation results of all these methods in an objective, quantitative manner, we compare the resulting segmentation masks using the method of [50], which uses a quality metric with higher perceptual meaning than simple spatial accuracy (i.e. the number of erroneous pixels). The segmentation mask errors are either missing foreground (MF) points, or added background (AB). The errors in the MF have a higher weight as the distance from the ground truth mask's border increases, whereas the errors in AB points stabilize after a certain distance from this border. This leads to the absolute Spatial Quality Measure (SQM) given by:

$$SQM = \sum_{d-1}^{D_{FG\max}} w_{MF}(d)\cdot Card(R_d \bigcap E^C) + \sum_{d-1}^{D_{FG\max}} w_{AB}(d)\cdot Card(R_d^C \bigcap E), \qquad (13)$$

where E is each estimated mask, R is the ground truth (reference) mask, $Card(S)$ is the cardinality operator for the set S, and the sets R_i are defined as:

$R_i = \left\{x \middle| x \in \Re, d(x, R^C) = i\right\}$.

The weighting factor w_{MF} is a linearly increasing function of the distance to the border of the ground truth mask, and w_{AB} is increasing until a distance *d*, after which it stabilizes [50]. The number of pixels in the SQM is higher when the estimated masks are visually unacceptable, so lower SQM values indicate better results.

The experiments were performed on the video sequences of Sec. V. Frames of size 720×576 were used in the Tennis Serve (63 frames), Tennis Hit (158 frames), and Soccer (100 frames). Frames of size 352×288 were used in the Pelikans sequence (200 frames), Table Tennis (100 frames), Surveillance (161 frames) and Surveillance-Hallway (100 frames). In Table I, we see that the SQM achieves its lowest values for the proposed, LRT-based method, and the next lowest for the EMD-based approach, in all videos except for the Tennis Hit sequence, where the EMD is lower. The CCF and color/motion clustering methods lead to higher values of SQM, as expected, since the corresponding segmentation results were not as good. The performance of the color cluster flow (CCF) approach [14] worsens as the non-rigidity of the moving objects increases. The algorithm of [15] leads to less reliable results, as a result of error propagation from the initial color segmentation stage, and the limitation of creating clusters with the same color and motion (a moving object might have different colors and different motions, e.g. a person in sports). The inferior performance of the latter two methods is expected, as they are based on clustering the motion or location of color homogeneous regions, which may be incorrectly segmented in the beginning, or which may be too many, due to over-segmentation. Finally, the proposed approaches based on the LRT and the EMD, have a similar computational time, of about 7 sec per video frame, on a dual core Pentium IV, for 720×576 frames, and about 4 sec for 352×288 frames. The methods of [14], [15] require more computations, of about 49 sec for 720×576 frames and 28 sec for 352×288 frames, as they are essentially applied on each sequence of homogeneous color clusters, over time, instead of being applied on the original video sequence only once.

TABLE I

QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS

Video	LRT method	EMD method	CCF	Color/motion clust.
Tennis Serve	**5.732**	6.673	9.443	8.982
Tennis Hit	6.321	**5.565**	9.522	9.233
Pelikans	**5.007**	6.343	5.788	7.752
Soccer	**5.211**	6.750	7.241	7.168
Table Tennis	**7.551**	8.689	10.011	9.878
Surveillance	**6.256**	8.572	9.399	10.432
Surveillance – Hallway	**5.141**	7.022	7.933	8.102

VII. Summary and Conclusions

In this paper we have presented a novel approach for the processing of video motion and color information, that can lead to the extraction of moving objects and the retrieval of the corresponding semantics. At a first stage, the motion estimates are processed statistically in order to extract activity areas for each frame. These activity areas are indicative of the events taking place. In order to fully describe the appearance of the scene and the moving objects, we also perform color processing of the video frames. Each activity and background area is segmented based on the colors present in it, and the background pixels are detected in the activity areas. This is achieved by comparing the colors of the active and static pixels via the EMD and a novel LRT based approach. Experiments show that the LRT method can give more reliable results than the EMD approach. Additionally, it results in a "likelihood ratio map" that shows with what probability each pixel belongs to a moving object or to the background, which can be used to tune the degree of segmentation desired in each case. Both methods outperform existing color-motion based approaches to segmentation, both in terms of segmentation quality and computational cost. Experiments show that the color information helps in the extraction of semantics from the video and that it effectively complements the motion information, by leading to accurate segmentation results.

Acknowledgments

This work was supported by the European Commission under contracts FP6-001765 aceMedia, FP6-027685 MESH and FP6-027026 K-Space.

References

[1] Ekin A. and A.M. Tekalp, "Shot type classification by dominant color for sports video segmentation and summarization," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on,* April 2003, vol. 3, pp. 173–176.

[2] Bertini M., Cucchiara R., Bimbo A., and A. Prati, "Semantic adaptation of sport videos with user-centred performance analysis," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 433 – 443, June 2006.

[3] Ekin A., Tekalp M., and Mehrotra R., "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796 – 807, July 2003.

[4] Yu X., Lai H.C., Liu S.X.F., and H.W. Leong, "A gridding hough transform for detecting the straight lines in sports video," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on,* July 2005.

[5] Intille S. and Bobick A., "Recognizing planned, multi-person action," *Comput. Vis. Image Understand.*, vol. 81, no. 3, pp. 414445, Mar. 2001.

[6] Tovinkere V. and Qian R. J., "Detecting semantic events in soccer games: Toward a complete solution," in *Proc. IEEE Int. Conf. Mult. Expo (ICME)*, Aug. 2001.

[7] Lie W., Lin T., and Hsia S., "Motion-based event detection and semantic classification for baseball sport videos," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, June 2004, pp. 1567 – 1570.

[8] Haritaoglu I., Harwood D., and Davis L. S., "W4: Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809830, Aug. 2000.

[9] Makris D. and Ellis T., “Learning semantic scene models from observing activity in visual surveillance,” *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 35, no. 3, pp. 397 – 408, June 2005.

[10] Fernyhough J. H., Cohn A. G., and Hogg D. C., *Generation of Semantic Regions from Image Sequences*, New York: Springer-Verlag, 1996.

[11] Koller-Meier E. B. and Van Gool L., “Modeling and recognition of human actions using a stochastic approach,” in *Proc. Eur. Workshop in Advanced Video-Based Surveillance Systems*, Sep. 2001, p. 1728.

[12] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, “Background and foreground modeling using nonparametric kernel density for visual surveillance,” in *Proceedings of the IEEE*, July 2002, vol. 90, p. 1151.

[13] B. Heisele, “Motion-based object detection and tracking in color image sequences,” in *Fourth Asian Conference on Computer Vision*, 2000, pp. 1028–1033.

[14] B. Heisele, U. Kressel, and W. Ritter, “Tracking non-rigid, moving objects based on color cluster flow,” in *Proc. ComputerVision and Pattern Recognition, CVPR 1997*, 1997, p. 253257.

[15] J. Guo, J. Kim, and C. Kuo, “New video object segmentation technique with color/motion information and boundary postprocessing,” *Applied Intelligence*, 1999.

[16] F. Dufaux and J. Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 497–501, March 2000.

[17] Bouguet J., “Pyramidal implementation of the lucas kanade feature tracker description of the algorithms,” in *OpenCV Documentation,Micro-Processor Research Labs, Intel Corporation*, 1999.

[18] J.L. Barron and R. Eagleson, “Recursive estimation of time-varying motion and structure parameters,” *Pattern Recognition*, vol. 29, no. 5, pp. 797–818, Dec. 1996.

[19] Kanade T. Lukas B., “An iterative image registration technique with an application to stereo vision,” in *International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

[20] B. Horn and B. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, pp. 185203, 1981.

[21] M.J. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise smooth flow fields.,” *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[22] F. Heitz and P. Bouthemy, “Multimodal estimation of discontinuous optical flow using markov random fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 12, pp. 1217–1232, 1993.

[23] B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills, “Recovering motion fields: An analysis of eight optical flow algorithms,” in *Proc. 1998 British Machine Vision Conference*, 1998.

[24] S. S. Beauchemin J. L.Barron, D. J. Fleet and T. A. Burkitt, “Performance of optical flow techniques,” in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR ’92, 1992 IEEE Computer Society Conference on*, June 1992, pp. 236–242.

[25] M. Sohling, M. Arigovindan, P. Hunziker, and M. Unser, “Multiresolution moment filters: Theory and applications,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 484–495, April 2004.

[26] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. 1, John Wiley & Sons, New York, 1966.

[27] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, second edition, 1987.

[28] H. V. Poor, *An Introduction to Signal Detection and Estimation*, Springer-Verlag, New York, second edition, 1994.

[29] G.B. Giannakis and M. K. Tsatsanis, “Time-domain tests for gaussianity and time-reversibility,” *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3460 – 3472, Dec. 1994.

[30] P. A. Delaney, “Signal detection using third-order moments,” *IEEE Transactions on Signal Processing*, vol. 13, no. 4, pp. 481–496, 1994.

[31] Hilbert D., *Color and Color Perception*, Cambridge University Press, 1987.

[32] Goldberger J. and Greenspan H., “Context-based segmentation of image sequences,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 463–468, March 2006.

[33] Cheng Y., “Mean shift, mode seeking and clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799.

[34] Hartigan J. A. and Wong M. A., “A K-means clustering algorithm,” *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[35] D. DeMenthon, “Spatio-temporal segmentation of video by hierarchical mean shift analysis,” in *Proc. of Statistical Methods in Video Processing Workshop*, 2002.

[36] Fukunaga K. and Hostetler L.D., “The estimation of the gradient of a density function, with applications in pattern recognition,” *IEEE Transactions on Information Theory*, vol. 21, pp. 32 – 40, 1975.

[37] Comaniciu V. and Meer P., “Kernel-based object tracking,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, May 2003.

[38] Comaniciu V. and Meer P., “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 603 – 619, May 2002.

[39] D. Comaniciu and P. Meer, “Robust analysis of feature spaces: Color image segmentation,” in *IEEE Conf. on Comp. Vis. and Pattern Recognition*, June 1997, pp. 750–755.

[40] G. Wyszecki and W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, New York: Wiley, 1982.

[41] Rubner Y., Tomasi C., and Guibas L. J., “The earth movers distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99 – 121, 2000.

[42] A. Elgammal, R. Duraiswami, and L. S. Davis, “Efficient non-parametric adaptive color modeling using fast gauss transform,” in *IEEE conference on Computer Vision and Pattern Recognition*, Dec. 2001.

[43] A. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction,” in *6th European Conference on Computer Vision*, June/July 2000.

[44] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking,” in *Computer Vision and Pattern Recognition, 1999. Proceedings CVPR ’99, 1999 IEEE Computer Society Conference on*, June 1999.

[45] Priebe C. E. and Marchette D. J., “Adaptive mixture density estimation,” *Pattern Recognition*, vol. 26, no. 5, pp. 771–785, 1993.

[46] M. B. Wilk and R. Gnanadisikan, “Probability plotting methods for the analysis of data,” *Biometrika*, , no. 55, pp. 1–17, 1968.

[47] Hanson K. and Wolf D., “Estimators for the Cauchy distribution,” in *Maximum Entropy and Bayesian Methods in Science and Engineering*, G. Heidbreder, Ed., pp. 255–263. Kluwer Academic, Dordrecht, 1996.

[48] G. R. Bradski and J. W. Davis, “Motion segmentation and pose recognition with motion history gradients,” *Journal Machine Vision and Applications*, vol. 13, no. 3, pp. 1432–1769, July 2002.

[49] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[50] P. Villegas, X. Marichal, and A. Salcedo, “Objective evaluation of segmentation masks in video sequences,” in *Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 1999*, 1999.