Fusion of Compound Queries with Multiple Modalities for Known Item Video Search

Ilias Gialampoukidis, Anastasia Moumtzidou, Stefanos Vrochidis and Ioannis Kompatsiaris
Information Technologies Institute - Centre for Research & Technology Hellas
Thessaloniki, Greece
Email: {heliasgj, moumtzid, stefanos, ikom}@iti.gr

***Abstract*—Multimedia collections are ubiquitous and very often contain hundreds of hours of video information. The retrieval of a particular scene of a video (Known Item Search) in a large collection is a difficult problem, considering the multimodal character of all video shots and the complexity of the query, either visual or textual. We tackle these challenges by fusing, first, multiple modalities in a nonlinear graph-based way for each subtopic of the query. In addition, we fuse the top retrieved video shots per sub-query to provide the final list of retrieved shots, which is then re-ranked using temporal information. The framework is evaluated in popular Known Item Search tasks in the context of video shot retrieval and provides the largest Mean Reciprocal Rank scores.**

I. Introduction

A plethora of multimedia collections is available in the Big Data era, containing images, videos, textual metadata and spatiotemporal information. Searching for a specific segment of a video is a challenging problem, even for personal video collections, due to the continuously increasing hours of video information recorded on a daily basis. This problem is known as the Known Item Search (KIS) task in annual competitions, such as the Video Browser Showdown[1] (VBS) and the TRECVID Ad-Hoc Video Search task. For example, in Fig. 1, the task is to find a part of a video of a yellow bus driving down winding road in front of a building with flags on roof and driving past geysers. The concepts of the description (geysers, bus, flags) formulate compound queries of multiple images, corresponding to each one of the description's concept. The main challenge in KIS task is to fuse different modalities, such as visual descriptors, concepts, color, temporal information and textual metadata that also appear in the query.

In video retrieval, the video is segmented into video shots that contain frames with similar content. Each shot comprises of frames, and the frame that is used for representing the whole shot is called keyframe. Thus, all features are computed on keyframe level. In video retrieval systems the query can be either visual or textual, aiming at the efficient retrieval of relevant shots. However, the queries are often complex, formulating compound queries of multiple and diverse concepts.

Compound queries in multimedia retrieval have appeared in [1], where the fusion is done at the feature level. For each image, that contains a face at a place, the face is cropped and FC7 visual descriptors are trained once on the face crop and once on the whole image. The result is two $L2$-normalized visual descriptors for the face and the place, respectively, which are then combined through an additional Support Vector Machine classification and a pairwise-minimum layer. We shall not restrict ourselves in two classes, such as faces and places, in order to be able to adapt the proposed method to two, three or more diverse and a priori unknown classes.

The novelty of this work is to propose a two-layer fusion method of visual descriptors, visual concepts and color features for combining multiple and diverse queries, where temporal information is also exploited. In contradiction to [2], we use compound queries and we also exploit the temporal order of video shots. We propose an integrated unifying approach that combines graph-based fusion of similarities at feature level, with late fusion at decision level. This methodology handles the complex nature of the multimodal query with multiple independent topics. We present a novel two-layer fusion video shot retrieval model with multimodal and multi-example compound queries, which is tested in popular Known Item Search tasks in video shot retrieval.

Our paper is organized as follows. Section II discusses related work in multimodal fusion for multimedia retrieval, Section III presents our proposed framework, Section IV contains the evaluation of our framework and, finally, Section V concludes the paper.

II. Related Work

Multimedia retrieval using multiple image queries has been tackled using several alternative methods in [3] and has been qualitatively evaluated in the TRECVID 2011 Known Item Search challenge [4]. The average query method (Joint-Avg), where the Bag-of-Words of all images (tf-idf scores) are averaged together has been compared against the Joint-SVM to retrieve shots from compound queries. In Joint-SVM a linear Support Vector Machine is used to learn a weight vector for visual words online at the feature level. The query set has the positive instances and a random set from the collection provides the negative instances. Visual words are weighted using an SVM-based supervisory layer, but the random choice of the negative class may affect the final results, since positive instances may also be included. A supervised late fusion approach is the Exemplar SVM (MQ-ESVM), which has originated in [5], that trains a separate linear SVM for each positive example and the final score is the maximal score.

[1]http://www.videobrowsershowdown.org/

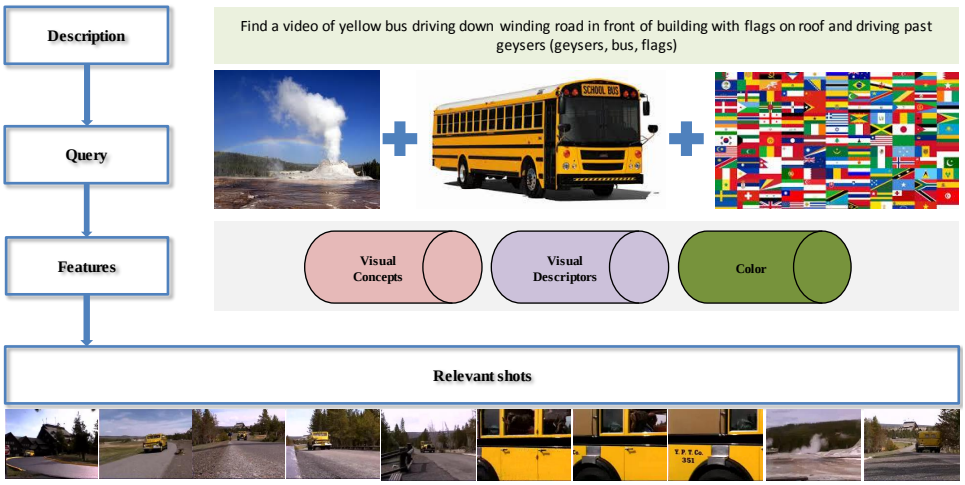

Fig. 1: Video Retrieval using compound queries

The combination of multiple retrieval tasks involve late fusion techniques of multiple rankings. SUMscore is a method for combining several ranked lists of retrieved results, in the context of video shot retrieval [6], where the authors conclude that features should first be fused separately for the visual examples and then these features' scores should be fused by adding the normalized relevance scores of the top searched results. Similarly, multiple queries are combined using the maximum or the average of individual scores obtained from each visual query [3] for unimodal search.

Rank-based late fusion approaches involve voting schemes such as the Condorcet Fusion [7], the Reciprocal Rank Fusion [8] and Borda fusion [9]. The late fusion problem becomes even more complex when multiple modalities (e.g. visual descriptors, concepts, timestamp, textual metadata, location) represent each video shot. A unifying model for unsupervised fusion of all similarities per modality has been presented in [10] and has been generalized to a non-linear fusion approach [2] that combines cross-media similarities with diffusion-based scores on the graph of items, in a non-linear but scalable way, for several modalities. Other methodologies for combining heterogeneous modalities involve Partial Least Squares [11], [12] and correlation matching, mapping multiple modalities to points in a common linear subspace. In [13] a video retrieval framework is proposed, which fuses textual and visual information in a non-linear way. We shall examine the performance of the late fusion approaches in multi-example and multimodal video shot retrieval.

Contrary to the aforementioned approaches, we provide a two-layer fusion framework that effectively fuses compound queries of diverse content, having multiple modalities, such as visual concepts and visual descriptors per modality. The first layer fuses multiple modalities per query using a graph-based non-linear fusion method. The second layer combines multiple ranked lists of retrieved multimodal objects, using also a temporal re-ranking stage to provide the final list of retrieved video shots, in response to a compound query.

III. Methodology

In KIS tasks, a visual or a textual description is provided to the system. The description usually has a list of concepts, formulating compound queries of multiple images, corresponding to each one of the description's concept. In the example of Fig. 1 the relevant video shot has a yellow bus but flags and geysers do not appear in all parts of the video segment. Each keyframe of the video segment is represented by visual descriptors, visual concepts and color features that need to be fused for each sub-query. We adopt the approach of [3], which involves formulating a corresponding group of multiple images by using the first retrieved image from Google images.

We firstly present the similarity fusion on multiple modalities and, then, the fusion of multiple ranking lists.

A. Background and Notation

The query $\mathbf{q}$ is formulated by U diverse images $q_u, u = 1, 2, \ldots, U$, each one having M modalities. Each image is retrieved from Google Images using separate keywords (e.g. geysers, bus, flags). The problem is to retrieve a ranked list r_u of multimedia items from a collection $\mathcal{M}$ of n items, in response to the query q_u and then combine the results for all $u = 1, 2, \ldots, U$. Graph-based models create graphs having nodes as multimedia items from the collection $\mathcal{M}$, and links weighted by transition probabilities from node κ to node λ. In the context of video retrieval, where given two similarity matrices S_1 and S_2 (one for each modality), a multimodal contextual similarity matrix C is computed [14]:

$$C = \beta_1 S_1 + \beta_2 S_2, \quad \beta_1 + \beta_2 = 1 \tag{1}$$

where $\beta_1, \beta_2 \in [0, 1]$. The matrix C is transformed to a row stochastic matrix P (all rows sum to one) when multiplied with the diagonal matrix D of size $n \times n$, with diagonal elements $d_{\kappa\kappa} = 1/\sum_{\lambda=1}^{n} d_{\kappa\lambda}$.

A unifying graph-based model has been proposed [10]:

$$\begin{aligned} x_{(i)} &\propto \mathbf{K}(x_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta_1 S_1 + \beta_2 S_2) + \gamma e \cdot s_1] \\ y_{(i)} &\propto \mathbf{K}(y_{(i-1)}, k) \cdot [(1-\gamma)D \cdot (\beta_1 S_2 + \beta_2 S_1) + \gamma e \cdot s_2] \end{aligned} \tag{2}$$

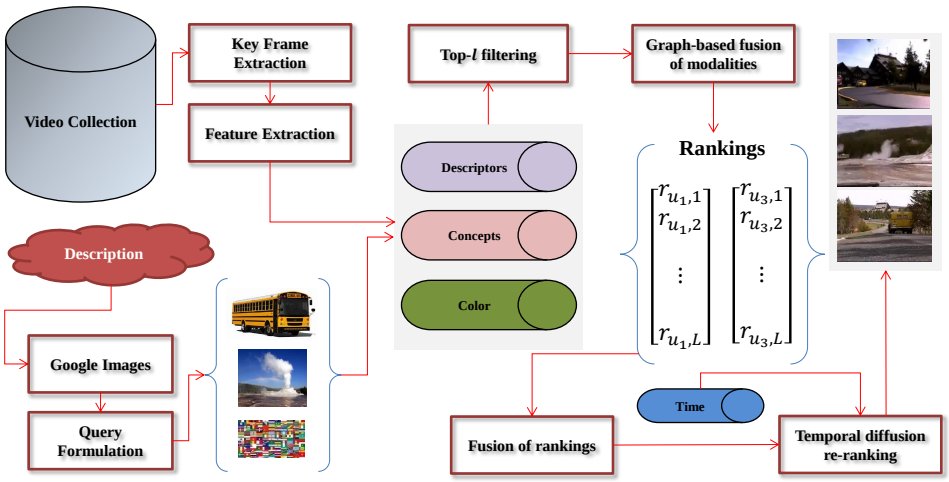

Fig. 2: Our Video Shot Retrieval framework

where $\mathbf{K}(.,k)$ is the operator that takes as input a vector and gives zero value to elements whose score is strictly lower than the k$^{\text{th}}$ highest value, e is the $l \times 1$ vector of ones, i is the number of iterations and the model sets $x_{(0)} = s_1$ and $y_{(0)} = s_2$, s_1 and s_2 are the query-based similarity vectors for the first and second modality, respectively. The number $l < n$ is fixed, usually set to $l = 1000$ and is defined as the number of "semantically filtered" multimedia items, with respect to the concepts' modality. After the initial filtering stage, l items are left, so the similarity matrices S_1 and S_2 are $l \times l$. The final relevance score vector is given by:

$$s^{graph}(q_u) = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 x_{(i)} + \alpha_4 y_{(i)} \qquad (3)$$

where $\sum_{m=1}^{4} \alpha_m = 1$ and $x_{(i)}, y_{(i)}$ are given by Eq. (2).

The first layer of our proposed two-layer fusion method combines multiple similarities per modality as follows:

B. Fusion of Multiple Modalities

Our framework fuses firstly multiple modalities and then fuses the results for each retrieved list of the compound query.

The model of Eq. (2) has been extended to multiple modalities [2], but not in the context of video retrieval. We modify the initial filtering stage, as shown in Fig. 2, filtering the key-frames by visual descriptors and not by concepts.

The model (2), in the case of three modalities, becomes:

$$\begin{aligned} x^1_{(i)} &\propto \mathbf{K}(x^1_{(i-1)}, k) \cdot [(1-\gamma_2-\gamma_3)\, P + \gamma_2 e \cdot s_2 + \gamma_3 e \cdot s_3] \\ x^2_{(i)} &\propto \mathbf{K}(x^2_{(i-1)}, k) \cdot [(1-\gamma_1-\gamma_3)\, P + \gamma_1 e \cdot s_1 + \gamma_3 e \cdot s_3] \\ x^3_{(i)} &\propto \mathbf{K}(x^3_{(i-1)}, k) \cdot [(1-\gamma_2-\gamma_1)\, P + \gamma_2 e \cdot s_2 + \gamma_1 e \cdot s_1] \end{aligned} \qquad (4)$$

where the contextual similarity matrix C, defined as

$$C = \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3, \qquad \beta_1 + \beta_2 + \beta_3 = 1 \qquad (5)$$

is row normalized to get the transition probabilities from one node to another: $p_{\kappa\lambda} = \frac{c_{\kappa\lambda}}{\sum_{\lambda=1}^{l} c_{\kappa\lambda}}$. The similarities among video shots for all modalities are calculated as [15], using $S_{\kappa\lambda} = 1 - \frac{E_{\kappa\lambda}}{max E_{\kappa\lambda}}$ formula, where $E_{\kappa\lambda}$ is the Euclidean distance between item κ and item λ.

The vectors of relevance scores $s^{nl\_graph}(q)$, in response to the query q, non-linearly combine the similarity vectors $s_m, m = 1,2,3$ and the vectors $x^m_{(i)}, m = 1,2,3$ in the case of graph-based and non-linear fusion:

$$s^{nl\_graph}(q_u) = s_1^{\alpha_1} + s_2^{\alpha_2} + s_3^{\alpha_3} + \alpha'_1 x^1_{(i)} + \alpha'_2 x^2_{(i)} + \alpha'_3 x^3_{(i)} \qquad (6)$$

The values of α_m, β_m, γ_m, $m = 1,2,3$ parameters are optimized following the methodology presented in [12], which involves keeping constant a set of parameters while examining the effect on the change of the others in the retrieval.

The model of Eq. (6) combines unsupervised multimodal fusion approaches, which are based on random walks, general graph diffusion processes and non-linear fusion of feature and cross-media similarities. The fusion of multiple modalities is one part of our proposed framework (first layer), and provides one list of retrieved video shots. Several lists need to be fused in the case of a compound query with multiple images, as shown in our framework (Fig. 2). In the following we present the second layer of our proposed two-layer fusion method, which fuses multiple rankings.

C. Fusion of Multiple Rankings and temporal re-ranking

In this work we extend the model of Eq. (6) to compound queries for video shot retrieval, based on the normalized scores of the top-l filtered video shots [6]. We denote by $\eta_u(shot)$ the similarity score between the sub-query u and each shot, as provided by Eq. (6): $\eta_u(shot) = s^{nl\_graph}(q_u, shot)$. On the top-$l$ retrieved shots, the scores η_u are normalized as follows:

$$\bar{\eta}_u(shot) = \frac{\eta_u(shot) - \min_{r(shot)<l} \eta_u(shot)}{\max_{r(shot)<l} \eta_u(shot) - \min_{r(shot)<l} \eta_u(shot)} \qquad (7)$$

The normalization of Eq. (7) provides scores in $[0,1]$ and only the graph-based non-linear fusion scores of the top-l

TABLE I: Mean Reciprocal Rank (MRR) results for the TRECVID 2011 and 2012 collections

	Fusion	Borda	Reciprocal	Condorcet	MINscore	SUMscore	MAXscore
2011	Concepts	0.0022 [9]	0.0021 [8]	0.0103 [7]	0.0014	0.0030 [6]	0.0043 [3]
	Descriptors	0.0042 [9]	0.0030 [8]	0.0028 [7]	0.0008	0.0043 [6]	0.0062 [3]
	Filtering_Concept	0.0024 [2]	0.0021 [2]	0.0011 [2]	0.0007 [2]	0.0032 [2]	0.0085 [2]
	Filtering_Descript	0.0036	0.0040	0.0030	0.0008	0.0040	**0.0090**
	Temp Re-ranking						**0.0100**
2012	Concepts	0.0465 [9]	0.0012 [8]	0.0082 [7]	0.0018	0.0509 [6]	0.0210 [3]
	Descriptors	0.0496 [9]	0.0072 [8]	0.0026 [7]	0.0014	0.0389 [6]	0.0428 [3]
	Filtering_Concept	0.0510 [2]	0.0007 [2]	0.0016 [2]	0.0018 [2]	0.0542 [2]	0.0742 [2]
	Filtering_Descript	0.0295	0.0115	0.0049	0.0014	0.0578	**0.0926**
	Temp Re-ranking						**0.0966**

retrieved results are taken into account. The normalized scores are then obtained per shot:

$$simil(shot) = \max_{u}\{\bar{\eta}_u(shot)\} \tag{8}$$

Temporal information is not available in [2], where the task is to retrieve text-image pairs, but is exploited to re-rank the top retrieved results. For each retrieved shot we check if the rank $r_u(shot)$ of the previous $(shot_{left})$ and the following $(shot_{right})$ shots belongs or not to the top 10% retrieved results with respect to the visual descriptors. In case this condition is true, the "temporal neighbor" is inserted at the position $r_u(shot) + 1$, shifting down the other results. The contribution of the temporal re-ranking stage is evaluated separately in the experimental comparison below.

IV. Experiments

In the following, we describe the datasets we used for evaluation, the features we extracted and the results.

A. Dataset Description

The datasets we have chosen for evaluation are annotated video collections from the two most recent KIS tasks of TRECVID, which are TRECVID 2011 and 2012 [4]. For the TRECVID 2011 a set of 25 topics were provided whereas for TRECVID 2012 a set of 24 topics were provided. Each topic was described adequately with a text description. From each textual description a set of Google images are extracted that to be used as multiple queries. For each one of the queries only one video segment is relevant, which is represented as a sequence of very few video shots. The task aims at finding the unique relevant video segment. We selected these datasets because the sub-query images are very diverse. Datasets with similar images per query-topic (e.g. pictures of Eiffel Tower) do not meet the assumptions of our problem.

B. Feature Extraction

For every keyframe, a set of features are extracted. Specifically, the features which are employed, are visual descriptors, visual concepts and color features. Visual descriptors are produced from deep convolutional neural networks (DCNNs). Specifically, a a GoogleNet [16] is trained on 5055 ImageNet concepts, and then, the output of the last pooling layer, with dimension 1024, is used as a global keyframe representation and thus as the visual descriptor. Regarding the visual concepts [17], the output of the GoogleNet was served as input to Support Vector Machine (SVN) classifiers. Specifically, one SVM is trained per concept, which results in 346 SVM classifiers, given that we used the 346 high level concepts from TRECVID. Finally, as color features we use the average euclidean distance to the pure red, green and blue.

C. Results

The comparison of the proposed framework is done using Mean Reciprocal Rank (MRR) scores and the results are presented in Table I. The evaluation under the first layer is done row-wise, where each row corresponds to multimedia retrieval using concepts, descriptors or all modalities considered (e.g. concepts, descriptors, color). The evaluation under the second layer is done column-wise, where each column evaluates the late fusion part of the framework [9], [3], [8], [7], [6].

Instead of using only one modality, we also compare against the similarity fusion and the "semantic filtering" of [2]. It should be noted that the results using solely color features are not presented because they are too low compared to the others given that method used is rather simple. The proposed two-layer framework without the temporal re-ranking stage is the combination of MAXscore with the filtering-by-descriptor stage (Filtering_Descript) and outperforms all baselines considered. We observe that the MAXscore performs better than the SUMscore [6], either using one or multiple modalities for comparison and that filtering by concept [2] underperforms when compared to the proposed filtering stage.

Moreover, for the TRECVID 2012 dataset, the results are promising, since the first relevant shot is retrieved, on average, on the top 10 positions out of 145,634 shots, bringing the relevant video shot to the first page in any video search engine.

V. Conclusion

We presented a two-layer multimodal multi-example video shot retrieval framework, based on graph-based fusion of several modalities and score-based fusion of multiple rankings. Our framework has been evaluated in two video collections of Known Item Search tasks, using the provided shots. In the future, we plan to add supervisory channels aiming at even better performance and model simplification.

Acknowledgment

This work was supported by the EC-funded projects beAWARE (H2020-700475) and V4Design (H2020-779962).

REFERENCES

[1] Y. Zhong, R. Arandjelović, and A. Zisserman, "Faces in places: Compound query retrieval," in *BMVC-27th British Machine Vision Conference*, 2016.

[2] I. Gialampoukidis, A. Moumtzidou, D. Liparas, S. Vrochidis, and I. Kompatsiaris, "A hybrid graph-based and non-linear late fusion approach for multimedia retrieval," in *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*. IEEE, 2016, pp. 1–6.

[3] R. Arandjelovic and A. Zisserman, "Multiple queries for large scale specific object retrieval." in *BMVC*, 2012, pp. 1–11.

[4] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. Smeaton, W. Kraaij, and G. Quénot, "Trecvid 2011-an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *TRECVID 2011-TREC Video Retrieval Evaluation Online*, 2011.

[5] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 89–96.

[6] K. Mc Donald and A. F. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *International Conference on Image and Video Retrieval*. Springer, 2005, pp. 61–70.

[7] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 538–548.

[8] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 758–759.

[9] J. A. Aslam and M. Montague, "Models for metasearch," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 276–284.

[10] J. Ah-Pine, G. Csurka, and S. Clinchant, "Unsupervised visual and textual information fusion in cbmir using graph-based methods," *ACM Transactions on Information Systems (TOIS)*, vol. 33, no. 2, p. 9, 2015.

[11] B. Siddiquie, B. White, A. Sharma, and L. S. Davis, "Multi-modal image retrieval for complex queries using small codes," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 321.

[12] I. Gialampoukidis, A. Moumtzidou, D. Liparas, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Multimedia retrieval based on non-linear graph-based fusion and partial least squares regression," *Multimedia Tools and Applications*, pp. 1–21, 2017.

[13] B. Safadi, M. Sahuguet, and B. Huet, "When textual and visual information join forces for multimedia retrieval," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 265.

[14] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 971–980.

[15] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 7, pp. 729–736, 1995.

[16] C. Szegedy, W. Liu, Y. Jia *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 00, pp. 1–9, 2015.

[17] F. Markatopoulou, V. Mezaris, and I. Patras, "Cascade of classifiers based on binary, non-binary and deep convolutional network descriptors for video concept detection," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1786–1790.