Temporal re-scoring vs. temporal descriptors for semantic indexing of videos

Abdelkader Hamadi, Philippe Mulhem, Georges Quénot

Temporal re-scoring vs. temporal descriptors for semantic indexing of videos

Abdelkader HAMADI [3] Philippe MULHEM [1,2] Georges QUÉNOT [1,2]
{Abdelkader.Hamadi, Philippe.Mulhem ,Georges.Quenot}@imag.fr
1: Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
2: CNRS, LIG, F-38000 Grenoble, France
3: Université de Lorraine

***Abstract*—The automated indexing of image and video is a difficult problem because of the "distance" between the arrays of numbers encoding these documents and the concepts (e.g. people, places, events or objects) with which we wish to annotate them. Methods exist for this but their results are far from satisfactory in terms of generality and accuracy. Existing methods typically use a single set of such examples and consider it as uniform. This is not optimal because the same concept may appear in various contexts and its appearance may be very different depending upon these contexts. The context has been widely used in the state of the art to treat various problems. However, the temporal context seems to be the most crucial and the most effective for the case of videos. In this paper, we present a comparative study between two methods exploiting the temporal context for semantic video indexing. The proposed approaches use temporal information that is derived from two different sources: low-level content and semantic information. Our experiments on TRECVID'12 collection showed interesting results that confirm the usefulness of the temporal context and demonstrate which of the two approaches is more effective.**

I. Introduction

The context helps to understand the meaning of a word and helps to disambiguate polysemous terms. Much research has exploited the advantage of this notion especially in information retrieval domain. Overall, it is recognized that the use of context enables the design of images analysis and understanding algorithms that are less complex and more robust [1]. For images and videos indexing and retrieval by visual concepts detection, this idea seems a priori valid. On the other hand, it is not always easy to have effective learners whatever the classes or concepts to deal with. Classic methods of semantic indexing of images and videos are not very effective because they manipulate multimedia samples and concepts separately and ignore contexts in which they appear. Some researchers tried to deal with this problematic. Indeed, several categories of context were considered: semantic [2], [3], [4], spatial [5], scale [6], temporal [7], [8], [9], [10], [11], [12], [13]. Videos have a characteristic that differentiates them from still images: the temporal aspect. Ignoring this feature leads to a loss of relevant information. In fact, the order of shots in a video gives a precise meaning to its content. So, the temporal structure of videos may be useful to bridge the semantic gap. On the other hand, to understand the content of a video, even a human needs to analyze information contained in some successive shots. This is due to the fact that the successive shots of a video are temporally, visually or semantically linked. Based on this idea, we assume that the presence of a concept in a given shot will increase the likelihood of its presence in certain shots that surround the concerned shot, especially the concepts that appear for a long period, as it is the case of events that usually last long. For example, if the concept "ski" appears in a video shot, its apparition is extended into several successive shots. In this same context, we will see the white color representing the "snow" throughout these successive shots. This proves that there is a dependency between the visual and/or the semantic contents of successive fragments of videos. Several researchers have attempted to use the temporal dimension for semantic video indexing and/or retrieval. However, the temporal information can be derived from different sources and exploited in different possible levels of an indexing system. We can cite: 1) before learning step, 2) in the learning process, and 3) in a post-processing step of the learning stage. In this paper, we compare two methods that concern the cases 1 and 3 for semantic indexing of videos.

II. Related works

The easiest way to use the temporal context is to consider it in a re-scoring step by post-processing the classification results. Safadi et al. [8], [9] used the notion of the temporal context for semantic indexing of videos by exploiting concepts detection scores in the neighboring shots. They achieved a very significant improvement of the baseline system. This good result can be explained by content dependency between successive shots that are locally homogeneous. We will detail this approach in section II-A. Temporal context can also be considered using temporal kernels. Qi et al. [11] propose a temporal kernel function. The problem with such methods is that they are not easy to implement and they usually require a long computation time and big data collections. Siersdorfer et al. [12] propose a video annotation approach based on context redundancy. Their idea assumes that most a video v_i labeled t_i includes a sequence of a video v_j labeled t_j, the likelihood that v_i will be labeled t_j increases. For videos annotation, Yuanning et al. [13] propose to consider the spatial and temporal contexts using a kernel function that takes into account the temporal correlation and the spatial correlation between concepts. For example, we may find a rule of the type: when a concept c_i is present in three successive shots, the concepts c_j and c_k co-occur in the last two successive shots. Experiments on TRECVid'05 and TRECVid'07 collections showed interesting results. Weng et al. [7] propose to integrate the semantic context (inter-concepts correlation) and the temporal context (the dependency between shots of a same video) for visual concept detection in videos. To create the contextual relationships between concepts, an algorithm similar to decision trees has been introduced. It is based on inter-concepts correlation calculated using the "chi-square" test on datasets updated progressively by subdividing the original collection. The temporal relationships are detected in the same manner as inter-concept relationships. So, a chronological route, in both directions (forward and backward) is performed. The operation continues until no shot with a significant correlation is found. In [14], the authors exploit the idea presented in [8], [9] and propose another approach that consists of generating a descriptor by performing an early fusion of high-level descriptors of shots belonging to a temporal window centered on the current shot. They achieved very interesting results and enhanced a good baseline system.

A. Temporal re-scoring (TRS)

We propose to use the idea of Safadi et al. [8], [9] by extending the target concept detection to the neighboring shots belonging to a window of size $2 \times w + 1$ shots instead of a single shot (w shots before the current shot and w shots after it). The new detection score of a concept c_i in the j^{th} shot e_j of a video v is given by the following formula:

$$F_{trs}(e_j, c_i) = \frac{1}{N} \times (F_{init}(e_j, c_i) + \sum_{k=-w, k\neq 0}^{w} F_{init}(e_k, c_i)) \tag{1}$$

where: e_k is the k^{th} shot of the video v. N is a normalization factor. $F_{init}(e_j, c_i)$ is the initial detection score of the concept c_i in the j^{th} shot of the video v. $F_{trs}(e_j, c_i)$ is the final detection score of the concept c_i in the j^{th} shot of the video v, obtained by a temporal re-scoring.

This fusion function is not unique. Indeed, many functions can be used, such as the one presented in [8], [9], where the new detection score of a concept c_i in a shot e is calculated as follows :

$$F_{trs}(e_j, c_i) = [F_{init}(e_j, c_i)]^{1-\gamma} \times [F_{temp}(e_j, c_i)]^{\gamma} \tag{2}$$

where γ is a parameter that controls the robustness of the re-scoring method, and $F_{temp}(e_j, c_i)$ is a global score that is calculated by merging the scores of the neighboring shots through a simple function such as arithmetic mean, geometric mean, min, max, etc. γ is tuned by cross validation on development set.

B. Temporal descriptors

TRS uses semantic information which is a set of detection scores of a visual concept in neighboring shots. Its performance is good when concept detectors have a reasonably good performance. However, if we take bad concept detectors we may be surprised by a degradation of performance. In fact, if such a detector provides erroneously the occurrence of a concept in shots s_{i-1} and s_{i+1}, it may improve the likelihood that the shot s_i contains the target concept while the latter does not belong to si. This error of temporal propagation of semantic information is due to errors made by concept detectors at the classification stage. We propose therefore, a method that uses the temporal context in a step that permits to avoid the problem of classification error propagation. It is possible to incorporate the temporal context in the descriptors extraction step in order to have features that consider themselves the temporal aspect. Descriptors of motion are a typical example for this case. One way to do this is to consider a temporal window when extracting descriptors instead of taking into account only the concerned shot. For example, to calculate a color histogram, we can count the occurrences of the gray levels values not only in one shot, but we extend the treatment to the neighboring shots. Similarly, for SIFT computing method, one can expand points of interest extraction into the neighboring segments, and we can also consider the neighborhood of a shot when counting the occurrences of visual words in bags of words computing. Although with such an idea we exploit temporal information, the temporal order is omitted. Indeed, after the histogram computing over a temporal window, it is not possible to distinguish the information corresponding to the current shot from those of the neighboring ones. Therefore, it becomes impossible to weight information of different sources. To remedy this problem we propose an approach that escapes these disadvantages and is much simpler to implement.

We propose to extract descriptors in a classic manner, from each segment and then make an early fusion by concatenating for each shot, the descriptors extracted from adjacent shots belonging to a temporal window of size equal to $(2 \times w + 1)$ successive shots and centered on the current one (w shots preceding and w shots following the current shot).

Although it is simple, this approach faces a major problem, which is the size of the resulting descriptor. Indeed, the final descriptor is $(2 \times w + 1)$ times larger than the feature vector extracted from a single shot in a classic situation. This may become critical especially for some types of big descriptors (e.g. Bags of opponent-SIFT). This affects the learning step which would be slow for large descriptors. It is therefore advisable to plan a dimensionality reduction step (e.g. PCA) to keep a reasonable size of descriptors. On the other hand, it is not recommended to consider a large temporal window for the same reasons of calculation performance. We propose in a second time, to weigh the temporal information so as to give more importance to the information extracted from the current shot. Weight values decrease as one moves away from the current shot. So, the more the shot is far from the center of the temporal window, the more its weight is low. Components values of each temporal fragment of the temporal descriptor are weighted by the weight value associated with it.

III. Experiments and Results

We tried at first to study the variation of relative gain provided by the temporal re-scoring regarding the baseline system performance. We used several baseline systems that follow the classic pipeline : "extraction/classification/fusion". You can refer to the work [14] for more details. These systems perform a late fusion of a set of concepts detectors trained on different types of low-level descriptors. The only difference in the experimental aspect of these systems is the choice and the number of fused descriptors and data collection used. Our evaluation was conducted on TRECVID 2010 and/or TRECVID 2012 collections. We used Mean Average Precision (MAP) calculated on the whole set of concepts considered as evaluation measure.

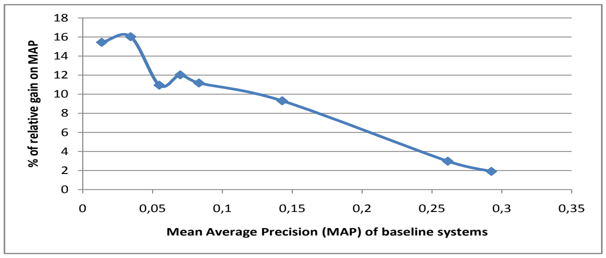

Fig. 1. % of relative gain on MAP achieved by TRS.

Figure 1 describes the variation of the percentage of the relative gain on MAP achieved by applying the temporal re-scoring on baseline systems. We note that the temporal re-scoring usually improves baseline system results. The percentage of relative gain on MAP varies between +2% and +16%. However, the performance of the studied approach may depend on the initial systems we want to improve. Indeed, we note that the relative gain is inversely proportional to the performance of the initial system. The more the system performance to improve is good, the more the gain obtained by the temporal re-scoring is low. This can be explained by the fact that it is easier to improve a system that is bad rather than a system with a good performance. On the other hand, we found that the gain obtained by the temporal re-scoring varies from one concept to another, but this approach improves in most cases concept detection in videos.

We compared in a second time the two approaches described above. To do this, we applied them on the same baseline system. We tested and evaluated our proposed approaches in the context of the semantic indexing task of TRECVid 2012 [15]. Annotations of 346 concepts were provided for the development set via the collaborative work [16]; and ground truth were provided for only 46 of them on the test corpus. We used for the evaluation measure the inferred average precision (infAP) on the 46 evaluated concepts.

We used a list of 10 variants of four types of descriptors: dense SIFT, lab, qw and vlat. You can refer to [17] for more details about these descriptors. We chose to use MSVM classifier/detector in the classification stage because of its good results [18] for video indexing. Before that, a preprocessing of descriptors was performed using a normalization (Power-law) and dimensionality reduction methods (ACP) [19]. After extracting the single descriptors of each video shot, we trained MSVM to generate a baseline model. Results of all single descriptors for each couple (concept, shot) are then fused by averaging the detection scores. We will call in the following this results: “Classic descriptors without TRS”. We considered it as a baseline system.

We then generated the temporal descriptors as described in section II-B. Because it was difficult to make experiments with a large temporal window, we considered in addition to the current shot the two shots surrounding it. We set weight values after a tunning stage on development data using descriptors of low sizes to facilitate the experimental process. We considered three cases (w_i is the weight value of the i^{th} shot): 1) $w_{i-1} = 1$, $w_i = 1$ et w_{i+1}=1, 2) $w_{i-1} = 0.05$, $w_i = 1$ et w_{i+1}=0.05, and 3) $w_{i-1} = 0.01$, $w_i = 1$ et w_{i+1}=0.01.

We generated for each of the ten single descriptors a temporal descriptor. So for each of these three cases, we obtained ten temporal descriptors. We then trained new MSVM detectors on the temporal descriptors. For each of the three cases and each pair (concept, shot), we combined the results of the ten temporal descriptors, by averaging the detection scores obtained from the classification step. We will call in the following this results: “Temporal descriptors without TRS”. We compare “Temporal descriptors” with TRS to study their complementarity. So we applied TRS on “Classic descriptors without TRS” and “Temporal descriptors without TRS”, respectively; to generate new results which we will call “Classic descriptors + TRS” and “Temporal descriptors + TRS”, respectively.

Results:

Systems			without TRS	+ TRS
Temporal descriptors	Weights	Prev. shot = 0.01 Current shot = 1 Next shot = 0.01	0.2330	0.2365
		Prev. shot = 0.05 Current shot = 1 Next shot = 0.05	0.2258	0.2285
		Prev. shot = 1 Current shot = 1 Next shot = 1	0.2052	0.2074
Classic descriptors			0.2312	0.2375

TABLE I. RESULTS (MAP) OF TEMPORAL DESCRIPTORS AND TEMPORAL RE-SCORING ON TRECVID 2012 COLLECTION.

Table I describes results obtained in terms of MAP obtained by the two approaches. The first column of MAP values presents the results obtained by a late fusion of the classic/temporal descriptors without applying TRS. The second one describes the results of applying the temporal re-scoring on the results of the classic/temporal descriptors late fusion. The last line concerns the late fusion of the classic descriptors. The first observation we can make is that the weighting step significantly affects the results. Indeed, we can see in table I that when the same weight value(= 1) is attributed to the different temporal segments we obtained a result worse than the one achieved by using classic descriptors. This leads to a system performance deterioration by decreasing the MAP from 0.2312 to 0.2052, which is equivalent to a relative difference of about -11.24%. Reducing weights of the neighboring shots leads to a performance improvement. The relative gain in this context is about +0.77%. We recall that for “Temporal descriptors”, we considered a temporal window of size equal to three successive shots. We expect better performance by adjusting and tuning the temporal window size and weights values. Moreover, we did not perform a more developed optimization of these parameters because of the experimental process heaviness, especially for large values of the temporal window and/or descriptors sizes. On the other hand, TRS improves the results in all cases, as we shown previously in figure 1. The relative gain varies here between +1% and +2.72%. Two other important remarks can be made. The first one is that the temporal re-scoring further improves the result of the classic descriptors fusion than the temporal ones, reaching a relative gain of about +2.72% ((0.2375-0.2312)/0.2312) and +0.77% ((0.2330-0.2312) /0.2312), respectively. The second remark is that TRS provides additional gain when it is applied to the fusion of temporal descriptors, reaching an additional relative gain of about +1.5 % ((0.2365-0.2330) /0.2330). This demonstrates the complementarity of the two methods that exploit the temporal context in different ways. In [14], TRS has deteriorated the results when it was combined with another method taking into account the temporal aspect. We explain this by the fact that both methods that were combined exploit temporal information of semantic type: detection scores. Indeed, this leads to a noise accumulation. But in our work, TRS uses temporal information of semantic type while “Temporal descriptors” exploits low-level information: descriptors. This explains somewhat why our two approaches are complementary and their combination improves the results.

An important question that we can ask is why the improvement achieved is not very high ? To answer this question we have verified the results per concept. We noticed that the gain achieved for some concepts (e.g. Motorcycle, Press_conference, etc) is very high and varies between +3,84%

and 43%. However, the described approaches fail and deteriorate the detection performance of some other concepts (e.g. Scene_Text, Walking_Running, etc). We explain this by the fact that some concepts appear for long periods (several consecutive shots) compared to some others, as is the case for: Airplane_Flying, Press_conference, Singing, etc. Considering the temporal context for such concepts proves a good idea. On the other hand, concepts for which our approaches fail mainly concern fixed objects or concepts that appear in scenes that contain many other different objects (e.g. Bicycling Appears with: road, trees, cars, houses, etc.). In addition, results also vary between descriptors. Therefore, with a simple fusion of their results, descriptors for which our approaches do not perform well affect negatively the overall performance.

Therefore, the gain achieved by our approaches is not low but they do not work well for some concepts and descriptors, and this affects the calculation of the overall performance measure (MAP). We believe that optimizing the fusion approach and choosing the concepts for which our approaches are appropriate could further improve the results.

IV. Conclusion and future work

We presented two methods exploiting the temporal context for semantic indexing of videos. The first one is a rescoring approach that merges concepts detection scores in the neighboring shots. This method demonstrated its effectiveness by improving several baseline systems reaching gains that are inversely proportional to the initial systems performance. The second approach incorporates the temporal context in descriptors extraction stage. To detect a concept c in a shot s_i, this approach combines the descriptors extracted from the neighboring shots by applying an early fusion of the descriptors belonging to a temporal window centered on s_i. Each of the temporal segments is weighted by a value, so as to give more importance to the shots that are closer to the current shot. Despite a few optimal use, this method has improved a baseline system, but it has been less effective than the temporal rescoring. However, we expect better performance by optimizing and tunning well the size of the temporal window and the weights values. Both approaches have shown the importance and usefulness of the temporal context for semantic video indexing and proved their complementary.

For the future work we will introduce other approaches in comparison and we will develop more deeply the study. To parallelize "Temporal descriptors" approach or studing its run on GPU architecture to better optimize it, take also part of our prospects.

Acknowledgements

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation. This work was supported in part by the French project VideoSense ANR-09-CORD-026 of the ANR. Experiments presented in this paper were carried out using the Grid'5000 experimental test bed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr). The authors wish to thank the participants of the IRIM (Indexation et Recherche d'Information Multimédia) group of the GDR-ISIS research network from CNRS for providing the descriptors used in these experiments.

References

[1] P. Brézillon, "Context in problem solving: A survey," *Knowl. Eng. Rev.*, vol. 14, no. 1, pp. 47–80, May 1999.

[2] M. R Naphade, I. V. Kozintsev, and T. S. Huang, "Factor graph framework for semantic video indexing," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 12, no. 1, pp. 40–52, Jan. 2002.

[3] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *ICCV*, 2007, pp. 1–8.

[4] Y. Qiu, G. Guan, Z. Wang, and D. Feng, "Improving news video annotation with semantic context," in *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, Dec 2010, pp. 214–219.

[5] J. Yuan, J. Li, and B. Zhang, "Exploiting spatial context constraints for automatic image region annotation," in *Proceedings of the 15th International Conference on Multimedia*, ser. MULTIMEDIA '07. New York, NY, USA: ACM, 2007, pp. 595–604.

[6] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vision*, vol. 53, no. 2, pp. 169–191, Jul. 2003.

[7] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *In Proceeding of the 16th ACM international conference on Multimedia, MM'08*, New York, NY, USA, 2008, pp. 71–80.

[8] B. Safadi and G. Quénot, "Re-ranking for Multimedia Indexing and Retrieval," in *ECIR 2011: 33rd European Conference on Information Retrieval*. Dublin, Ireland: Springer, Apr. 2011, pp. 708–711.

[9] B. Safadi and G. Quénot, "Re-ranking by Local Re-scoring for Video Indexing and Retrieval," in *20th ACM Conference on Information and Knowledge Management (CIKM)*, Glasgow, Scotland, Oct. 2011, pp. 2081–2084.

[10] A. Hamadi, G. Quénot, and P. Mulhem, "Two-layers re-ranking approach based on contextual information for visual concepts detection in videos," in *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, 2012, pp. 1–6.

[11] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H.-J. Zhang, "Correlative multilabel video annotation with temporal kernels," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 5, no. 1, pp. 3:1–3:27, Oct. 2008.

[12] S. Siersdorfer, J. San Pedro, and M. Sanderson, "Automatic video tagging using content redundancy," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 395–402.

[13] Y. Li, Y. Tian, L.-Y. Duan, J. Yang, T. Huang, and W. Gao, "Sequence multi-labeling: A unified video annotation scheme with spatial and temporal context," *Multimedia, IEEE Transactions on*, vol. 12, no. 8, pp. 814–828, Dec 2010.

[14] A. Hamadi, P. Mulhem, and G. Quénot, "Extended conceptual feedback for semantic multimedia indexing," *Multimedia Tools Appl.*, vol. 74, no. 4, pp. 1225–1248, 2015.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A. F. Smeaton, and G. Quénot, "TRECVID 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[16] S. Ayache and G. Quénot, "Video Corpus Annotation using Active Learning," in *European Conference on Information Retrieval (ECIR)*, Glasgow, Scotland, Mar. 2008, pp. 187–198.

[17] N. Ballas, B. Labbé, A. Shabou, H. Le Borgne, P. Gosselin, M. Redi, B. Merialdo, H. Jégou, J. Delhumeau, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, F. Thollard, N. Derbas, G. Quénot, H. Bredin, M. Cord, B. Gao, C. Zhu, Y. tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benot, P. Lambert, T. Strat, J. Razik, S. Paris, H. Glotin, T. Ngoc Trung, D. Petrovska Delacrétaz, G. Chollet, A. Stoian, and M. Crucianu, "IRIM at TRECVID 2012: Semantic Indexing and Instance Search," in *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, 2012.

[18] B. Safadi and G. Quénot, "Evaluations of multi-learners approaches for concepts indexing in video documents," in *RIAO*, Paris, France, 2010.

[19] B. Safadi and G. Quénot, "Descriptor optimization for multimedia indexing and retrieval," in *Proc. of Content Based Multimedia Ingexing (CBMI) Workshop*, Veszprém, Hungary, June 2013.