Multimodal Fusion for Video Copy Detection*

Xavier Anguera
Telefonica Research
Torre Telefonica Diagonal 00
08019 Barcelona, Spain
xanguera@tid.es

Juan Manuel Barrios
PRISMA Research Group
Department of Computer
Science, University of Chile
jbarrios@dcc.uchile.cl

Tomasz Adamek
Telefonica Research
Torre Telefonica Diagonal 00
08019 Barcelona, Spain
tomasz@tid.es

Nuria Oliver
Telefonica Research
Torre Telefonica Diagonal 00
08019 Barcelona, Spain
nuriao@tid.es

ABSTRACT

Content-based video copy detection algorithms (CBCD) focus on detecting video segments that are identical or transformed versions of segments in a known video. In recent years some systems have proposed the combination of orthogonal modalities (e.g. derived from audio and video) to improve detection performance, although not always achieving consistent results. In this paper we propose a fusion algorithm that is able to combine as many modalities as available at the decision level. The algorithm is based on the weighted sum of the normalized scores, which are modified depending on how well they rank in each modality. This leads to a virtually parameter-free fusion algorithm. We performed several tests using 2010 TRECVID VCD datasets and obtain up to 46% relative improvement in min-NDCR while also improving the F1 metric on the fused results in comparison to just using the best single modality.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms

Keywords

Content-Based Video Copy Detection, Multimodal, Fusion

1. INTRODUCTION

Content-based video copy detection systems aim at finding video segments that are identical or transformed versions of segments in an known video. Joly et al. [4] propose a definition of video copy based on a subjective notion of tolerated transformations. A tolerated transformation is a function that creates a new version of a document where the original document "remains recognizable". CBCD systems perform the detection by processing visual and/or audio content of videos, ignoring any metadata and avoiding to embed watermarks into the original videos.

Traditionally, CBVCD systems have relied only on the visual information in the videos, driving research towards new features that allowed for scalable systems and fast retrieval. In recent years systems started incorporating features derived from the audio modality. Through the fusion of multimodal information, CBVCD systems can obtain improved performance levels and be more robust to errors (e.g. one of the modalities is severely damaged or missing in the video copy). An important step in the combination of multiple input streams (which can be more than two, including several system results from audio and from video) is how their fusion is performed. Typically, there are significant differences in how the different modalities internally produce their results. Hence, most systems perform the fusion at the decision level by combining the individual outputs.

In this paper we propose a novel weighted-sum decision-level multimodal fusion system that is able to effectively combine an unlimited number of monomodal outputs, consisting of lists of the best matches sorted by scores. The main strengths of the proposed approach are three-fold: (1) The use of a *flooring bias* on the individual output scores, combined with an area-preserving normalization, ensures that all individual modalities have an equivalent impact in the final scores, which is dependent on the ratio between the different matching segments; (2) the *rank* of each individual result is used as extra information in the fusion; (3) as a side result, the weights in the final sum of scores can be set to trivial values with no significant decrease in performance, therefore turning our system into virtually *parameter-free*.

2. MULTIMODAL FUSION APPROACHES

To the present day, most of the proposed combinations consist on the fusion at decision level. They can be classified into two, rather fuzzy, classes: rule-based fusion and weighted sum of the output scores.

On the one hand, rule-based fusion approaches apply some

*Area chair: Kiyoharu Aizawa

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

heuristic rules to choose which of the results from each modality should be considered in the final results. Saracoglu et al. [11] return the single match with highest score between both modalities. Liang et al. [6] combine the multiple outputs from each modality either through the union (less restrictive) or the intersection (more restrictive and less prone to false alarms) of all results. Mukai et al. [9] obtain very good performance by considering the overlap between video matches in both modalities and giving priority to the audio results when selecting the output results.

On the other hand, many systems do a weighted sum of the scores obtained from each modality. Like in [6], Le et al. [5] apply the union/intersection of the two modalities, and then in the overlapping segments they apply a weighted sum of the scores with manually set weights (emphasizing the audio scores). One important step that is many times overlooked by the systems is the dynamic range of scores obtained by each modality, which makes direct comparison of system outputs difficult. Among the few systems that describe the scores normalization being applied are Narsev et al. [10] who normalize each modality by the median scores obtained on a development set, and Uchida et al. [13] who normalize scores by the second-best score in the resulting list of matches and then apply some rules for score boosting depending on how many modalities contain the same match.

Probably the closest work to our proposal is from Jegou et al. [3] who perform an initial normalization of the features and then combine them using a weighted sum with equal weight for each modality. In addition, they include a score weighting dependent on the ratio of each match with the best match in that modality. Our approach is different in that our normalization already takes into account the best-matching score and in addition we consider the position of the match within the ranked list in each modality.

3. MULTIMODAL VIDEO COPY DETECTION SYSTEM

In this section we briefly describe the parallel audio and video processing modules that run at the core of our system to obtain the outputs that are later fused with the proposed multimodal fusion algorithm. Each of the modules finds, for each given query, a list of up to N potential copies within a reference database, and returns the start-end times for both query and reference videos in the matching segment, and a score. Note that at this point the output range of the scores is not enforced to match in all cases, which will be dealt with by the fusion preprocessing step. The fusion system presented in this paper has been tested using the output of three individual modules: a global video module, a local video module and a local audio module. These modules were developed independently by two TRECVID 2010 participating teams and obtained good results individually.

3.1 Visual Global Features Module

The visual global features module was developed by PRISMA Research Group at the University of Chile [1]. It defines a distance between video segments as a weighted combination of global descriptors and uses a pivot-based indexing structure to perform approximate searches. It is divided in five tasks: Preprocessing (which tries to minimize the effect of visual transformations by normalizing video frames), Video Segmentation (which partitions every video into small segments of nearly $\frac{1}{3}$ seconds length), Feature Extraction (which represents each segment with the average Edge Histogram [7] and the average Gray Histogram), Approximate Search (which for every query segment performs an approximate k-NN search retrieving the most similar reference segments), and Copy Localization (which looks for chains of similar reference segments, and returns a list copy candidates each one with a location and a score).

3.2 Visual Local Features Module

The local visual features processing module was developed at Telefonica Research [14]. It compares all query keyframes with all reference keyframes using a state-of-the-art image retrieval engine relying on local features [8] and then combines the obtained ranked lists of matched keyframes into copied video segments by performing a temporal consistency post-processing. The module is divided in three tasks: Feature Extraction (which samples every video with one frame per second and extracts novel local features called DART [8]), Keyframe Matching (by using hierarchical dictionaries of visual words and inverted files to locate matching frames, which are then refined through a spatial verification stage), and Temporal Consistency (which computes the time differences between matches, and returns a list copy candidates each one with a location and a score).

3.3 Audio Local Features Module

The audio local features module was also developed at Telefonica Research [14]. Our implementation is similar to Philips [2]. First, the feature extraction computes the short-time FFT of the acoustic data every 10ms, converts the frequency bins into Mel scale, and codifies the first bands with a 15-bit fingerprint. Then we insert the data into a hash table and search for exact matches between query and reference information. Finally, a temporal consistency step (similar to the video one) is used to obtain the matching segments between videos.

4. MULTIMODAL FUSION ALGORITHM

Once all individual modules have computed their N_k-best matching results for any given query, the fusion module merges all these results into a single output ranking in order to a) reduce the false alarm rate from matches present in any of the outputs, and b) reduce the miss rate of any individual modality. The final result of the fusion algorithm is a ranked list of the N-best overall matches, together with their final score. Such scores are normalized to the range [0,1] to make it easier to later apply a copy decision threshold, regardless of how many (and which) modalities have been merged. The output of N_k-best results is convenient to compute optimal score curves, while in a real-case scenario we would only output those matches (if any) exceeding a given application-specific decision threshold.

The fusion algorithm performs two main tasks. First, it finds and merges the matching results coming from different modalities when they are in overlap. Then, it computes a final matching score for all segments, both the merged ones and also those coming from only one modality. When doing do, we take into account both the similarity score and the ranking obtained by the videos in each modality. Next we describe in detail each of the steps involved in the fusion.

4.1 Scores Preprocessing and Normalization

The inputs to the algorithm are the lists of N_k-best reference video matches from the available K input modalities for a given query, ordered by their matching score $S_k(r)$ $|r\epsilon\{1 \dots N_k\}$, $k\epsilon\{1 \dots K\}$. Note that the dynamic range of the scores for every modality does not need to be the same, but all need to represent a similarity, where value 0 represents an exact match. In the case that any of the input modalities is not able to return the same number of matches as the other modalities (*i.e.* $N_k < N$) it causes a potential problem as the following normalization steps would artificially emphasize these modalities more than the others. To void this problem we apply a preprocessing step, that we call N-best match flooring, which consists of adding a small value α to all scores, and extending the number of results to have $N_k = N$ output scores, as shown in Eq. 1. The value of α can be fixed for all modalities or set dynamically depending on their statistics. In our case we set $\alpha = 0.1$ for all cases.

$$S'_k(r) = \begin{cases} S_k(r) + \alpha & 1 \leq r \leq N_k \\ \alpha & N_k < r \leq N \end{cases} \quad (1)$$

Next, the matching scores of each modality k, $S'_k(r)$, are independently L1-normalized in order to make them comparable with each other. For each score in modality k we normalize it as $\hat{S}_k(r) = \frac{S'_k(r)}{\sum_{j=1}^{N} S'_k(r)}$. Note that the underlying distribution of scores within each modality remains intact with such normalization.

4.2 Fusion of Normalized Scores

After preprocessing all scores we fuse them by considering their ranking r within each modality, their normalized scores and the temporal limits.The parameters associated with each matching segment $c_k(r)$ in each of the computed modalities are: $c_k(r) = \{B_k^Q(r), E_k^Q(r), B_k^R(r), E_k^R(r), \hat{S}_k(r), I_k(r)\}$, where $B_k^Q(r) \dots E_k^R(r)$ are the start-end times of the matching segments both for query and reference videos, $\hat{S}_k(r)$ is the matching score and $I_k(r)$ is the ID of the reference video the segment matches with.

Given all matching segments found in the different modalities, in this step we want to create a set of L fused segments $C = \{c_1 \dots c_L\}$ containing both new segments created from the overlap of original segments in individual modalities and the rest of original matching segments that did not overlap with others. For any two matching segments $c_{k_1}(r_1)$ and $c_{k_2}(r_2)$ (or alternatively between a matching segment and a partially fused segment) we determine they are in overlap if $I_{k_1}(r_1) = I_{k_2}(r_2)$ and

$$\frac{\min\{E_k^Q(r), E_k^R(r)\} - \max\{B_k^Q(r), B_k^R(r)\}}{\max\{E_k^Q(r), E_k^R(r)\} - \min\{B_k^Q(r), B_k^R(r)\}} > 0.5$$

Note that the way that matches are retrieved in every individual modality makes it impossible to obtain multiple matches from the same modality in the same final segment. When two segments are in overlap we fuse their segment boundaries (both for query and reference) selecting as start time the minimum between all segments' start times, and as end time the maximum between all end times. Finally, given all matching segments $c_k(r)$ that have been fused into a c_l, we obtain the final score of c_l as

$$S(c_l) = \frac{\sum_{c_k(r)\epsilon c_l} W_k \cdot \frac{N_k - r + 1}{N_k} \cdot \hat{S}_k(r)}{\sum_{k=1}^{K} (W_k \cdot \hat{S}_k(1))} \quad (2)$$

where the ranking r of the segment within a given modality k affects the final score through the term $\frac{N_k - r + 1}{N_k}$, which is 1 for the best match and $\frac{1}{N_k}$ for the worst. Additionally, the term W_k is an optional weight parameter to manually emphasize some modalities versus others in the final score. As will be seen in the evaluation, we only use this parameter to balance the impact of the audio versus the video modalities. Note that the scores are normalized by the sum of all best matching scores for each modality, $\hat{S}_k[1]$, such that fused segments lacking matching segments for some of the modalities will incur in some penalty, and also all scores will be in the $[0, 1]$ range.

Once all $S(c_l)$ have been computed, they are ranked and the matching clusters with the N-best scores are returned, discarding the rest. Alternatively, an application-dependent threshold could be used to output the matches (if any) that exceed its value.

5. EXPERIMENTS

We tested the suitability of the proposed algorithm using the content-based copy detection dataset used in the 2010 NIST TRECVID benchmarking campaign [12]. For a set of 11,256 query videos the task consists on finding whether (and which) portions appear in any in any of the 11,524 videos in the reference dataset. In the creation of the queries 8 possible video transformations and 7 possible audio transformations from the reference videos have been used.

To evaluate the performance we use the The Normalized Detection Cost Rate (NDCR) and the F1 score like in the TRECVID evaluations. Perfect results would obtain NDCR=0 and F1=1. In addition, we also report the percentage of videos with correct matches appearing within the N-best results, independently of their score. Even if some of these matches might have a low score and never be chosen without incurring in a very high false alarm rate, this metric gives an upper bound of the number of true positives that each system can achieve.

5.1 Experimental Results

We performed experiments both in the individual modalities and in all possible combinations of two and three modalities. Given that we fuse two video-based modalities with one audio modality, when combining all three modalities we decided to balance the contribution of audio and video by setting $W_k = 0.5$ for the video modalities. Otherwise we set $W_k = 1$. In all the experiments we set $N = 20$.

Table 1: *NDCR and F1 results for the combination of multiple modalities*

Output conmbination	Min NDCR	Opt F1	%hits
Local Audio (LA)	0.928	0.952	75.5%
Local Video (LV)	0.928	0.930	84.7%
Global Video (GV)	0.715	0.893	65.5%
LA + GV	0.664	0.954	93.3%
LV + GV	0.824	0.949	90.8%
LA + LV	**0.590**	0.950	91.9%
LA + LV + GV	0.598	**0.964**	96.9%

Table 1 presents the minimum NDCR, optimum F1 scores and the percentage of hits for all possible combinations of input modalities being considered. The first block of results show the scores for each modality output independently, namely for the local audio system (LA), the local video system (LV) and the global video system (GV). In the second and third blocks we do the proposed fusion in groups of two and three modalities, respectively.

In terms of the NDCR and F1 metrics (second and third columns on the Table), the global video system obtains significantly better NDCR results than any of the local features systems, but its F1 score is the lowest. This indicates that the matching start-end times for this modality are not as accurately found as in the other modalities. In the second block we observe that any combination of *audio and video* features yields a significant improvement in NDCR score over any individual modality. However, the NDCR score of the local and global video combination is worse than the best individual video result (GV). This result shows the correlation of the two approaches, even if they focus on different characteristics of the video modality. Still, the improvement in F1 score obtained when combining LV+GV makes it worthwhile to use the fusion. Similarly, the F1 score is already very good when using LA alone and remains similar when combining these with any of the video features.

Finally, the combination of all individual systems obtains a slightly worse NDCR score than the best combination of two modalities (LA + LV), while the F1 score obtains an important improvement.

In terms of (*%hits*), they progressively increases as more modalities are taken into account. This indicates that the fusion algorithm can incorporate and give higher scores to the useful information in each individual modality.

Table 2: *Overlap between matches in single modalities and LA+LV+GV*

	LA	LV	GV
total matches	5664	6349	4914
Shared with LA+LV+GV	99.9%	100%	99.8%
Only in LA+LV+GV	28.5%	14.4%	48.1%

Next we analyze how well the real matches get transferred from the N_k-best results on the individual modalities to the final N-best results, given that the final number of results is $\frac{1}{K}$ the size of the sum of results of all individual modalities. Table 2 shows the matches overlap between correct matches found by each individual modality (found within the N-best returned results) and by the combination of all three modalities. The first line in the table shows the number of total correct matches that each individual modality returns (local video is the one that returns the biggest number). The second line shows how many of these individual matches are present in the results from the fusion algorithm. We see how in all cases most matches found in the individual modalities are contained in the fused results. This ensures that our fusion system does incorporate correctly the information present in the individual modalities. Finally, for completeness purposes, we show the percentage of matches in the fusion results that do not appear in the individual modalities.

6. CONCLUSIONS AND FUTURE WORK

In this paper we propose a fusion algorithm for the combination of multiple modalities at the decision level. The final list of possible matches is obtained through the weighted sum of the normalized scores for each modality, modified depending on how well they rank in each modality. We perform extensive tests with the TRECVID 2010 dataset and show that the fused results improve by more than 46% relative the min-NDCR results. We also observe that the results are more stable in the multimodal system than in any of the multimodal modules. Finally, we show that the F1 scores improve when considering multiple overlapping segments. Future work will include testing the proposed approach with more monomodal processing outputs to see whether our conclusions hold for more than 3 modalities, and to further investigate the role that different modalities and features play on the CBCD task.

7. REFERENCES

[1] J. M. Barrios and B. Bustos. Content-based video copy detection: PRISMA at trecvid 2010. In *Proc. NIST-TRECVID Workshop*, 2010.

[2] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proc. ISMIR*, 2002.

[3] H. Jegou, M. Douze, G. Gravier, C. Schmid, and P. Gros. INRIA LEAR-TEXMEX: Video copy detection task. In *Proc. NIST-TRECVID Workshop*, 2010.

[4] A. Joly, O. Buisson, and C. Frélicot. Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Trans. on Multimedia*, 9(2):293–306, 2007.

[5] D.-D. Le, S. Poullot, M. Crucianu, X. Wu, M. Nett, M. E. Houle, and S. Satoh. National institute of informatics, japan at TRECVID 2009. In *Proc. NIST-TRECVID Workshop*, 2009.

[6] Y. Liang, B. Cao, J. Li, C. Zhu, Y. Zhang, C. Tan, G. Chen, C. Sun, J. Yuan, M. Xu, and B. Zhang. THU-ING at TRECVID 2009. In *Proc. NIST-TRECVID Workshop*, 2009.

[7] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. on Circuits and Systems for Video Tech.*, 11(6):703–715, 2001.

[8] D. Marimon, A. Bonnin, T. Adamek, and R. Gimeno. DARTs: Efficient scale-space extraction of daisy keypoints. In *Proc. CVPR*, 2010.

[9] R. Mukai, T. Kurozumi, K. Hiramatsu, T. Kawanishi, H. Nagano, and K. Kashiro. NTT communications science laboratories at TRECVID 2010 content-based copy detection. In *Proc. NIST-TRECVID Workshop*, 2010.

[10] A. Narsev, S. Bao, J. Chang, M. Hill, M. Merler, J. R. Smith, D. Wang, L. Xie, R. Yan, and Y. Zhang. IBM research TRECVID-2009 video retrieval system. In *Proc. NIST-TRECVID Workshop*, 2009.

[11] A. Saracoglu, E. Esen, T. Ates, B. O. Acar, U. Zubari, E. C. Ozan, E. Ozalp, A. A. Alatan, and T. Ciloglu. Content based copy detection with coarse audio-visual fingerprints. In *Proc. CBMI*, pages 213–218, June 2009.

[12] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[13] Y. Uchida, S. Sakazawa, M. Agrawal, and M. Akbacak. KDDI labs and SRI international at trecvid 2010: Content-based copy detection. In *Proc. NIST-TRECVID Workshop*, 2010.

[14] E. Younessian, X. Anguera, T. Adamek, N. Oliver, and D. Marimon. Telefonica research at trecvid 2010 content-based copy detection. In *Proc. NIST-TRECVID Workshop*, 2010.