Telefonica Research at TRECVID 2011 Content-Based Copy Detection

Xavier Anguera[1] Tomasz Adamek[1], Daru Xu[2] and Juan Manuel Barrios[3]
[1]Telefonica Research,
Torre Telefonica-Diagonal 00, 08019 Barcelona, Spain.
[2]Ming-Hsieh Department of Electrical Engineering,
University of Southern California, USA.
[3]PRISMA Research Group,
Department of Computer Science, University of Chile, Chile.
{xanguera, tomasz}@tid.es

***Abstract*—This notebook paper summarizes the algorithms behind Telefonica Research participation in the NIST-TRECVID 2011 evaluation on the Video Copy Detection task. This year we have focused on 1) Improving the image-based matching system to better process video files; 2) implemented and tested a novel audio local fingerprint; and 3) improved the multimodality fusion algorithm from last year.**

For this year we have submitted 4 runs in total, whose main characteristics are described below:

- **TID.m.[BALANCED/NOFA].multimodal: These correspond to our main submissions, both for the no false alarm and balanced profiles. They are based on the fusion between the local audio and local video monomodal systems.**
- **TID.m.BALANCED.mask: This submission is based on the monomodal audio-based system, which this year uses a novel audio fingerprint called MASK.**
- **TID.m.BALANCED.joint: This submission is the fusion (at decision level) from our two monomodal system outputs with the output from the PRISMA group video-only system. This submission resulted in our best results for the evaluation.**

Over all, we are very pleased with the results for this year's evaluation. On the one hand, our video-based system is reaching maturity, using local image descriptors (DART) developed by Telefonica. On the other hand, we have developed and applied to the evaluation novel audio local features called MASK. Even though we did not spend much time tuning the new feature to the Trecvid copy detection datasets, we are very please with its results. In addition, we have improved the fusion algorithm from last year and have shown that it does work well to fuse results from multiple outputs, improving on the results obtained by either one of our systems and those from the PRISMA submission.

I. Introduction

The final goal in the video copy detection task, within the NIST-TRECVID evaluation campaign [1], is to locate segments within given query videos that occur, with possible transformations, in a given reference video collection. Applied transformations can be inherent to the general video creation process (like encoding artifacts, video quality changing, etc.) or more complex transformations, which manipulate video content or its orientation (e.g., flipping, frame dropping, cropping, insertion of text/patterns like fixed banners or logos, etc.). In the audio track the possible transformations go from mild (e.g. MP3 transcoding) to very severe (overlapping speech plus various companding effects.).

Video copy detection can play an essential role in many real-live applications, for example search result redundancy removal, copyright control, business intelligence, advertisement tracking, law enforcement investigations, etc. In addition to the conventional method of watermarking, content-based copy detection is considered an alternative solution for video copy detection. In the watermarking approach irreversible information (i.e. watermarks) is embedded in the original video stream and is used to determine if a video has been copied from another video. One limitation of this approach is that the distributed videos should have been watermarked in the source, which adds an extra post-processing step for production companies or individuals, which not always can be done, as many times we do not have access to the source video. On the other hand, through content-based approaches, a set of content-based features are extracted from the video and are utilized to locate copied segments of the query video in a reference video dataset. It is said that in the content-based approach the content itself acts as the watermark. This approach is still a challenging topic because of various types of transformations applied and computational issues, although research on the area is progressing steadily.

In general, the main challenges in video copy detection are: 1) scalability to deal with growing amounts of multimedia data; 2) speed of new media indexing into the reference database; 3) speed of retrieval of plausible copies, given a query video; 4) effective usage of the audio+video information available; and 5) effectiveness on finding the correct duplicates while reducing the number of false alarms.

In NIST-TRECVID video copy detection task, the focal point is to evaluate content-based approaches. For the second year in a row, this year's evaluation is focused on finding copy segments from a database that resembles videos we can find in Internet sharing sites. In total there are over 400 hours of reference videos and a total of over 11K queries have to

[2]Mr. Xu participated in the evaluation during an internship at Telefonica Research.

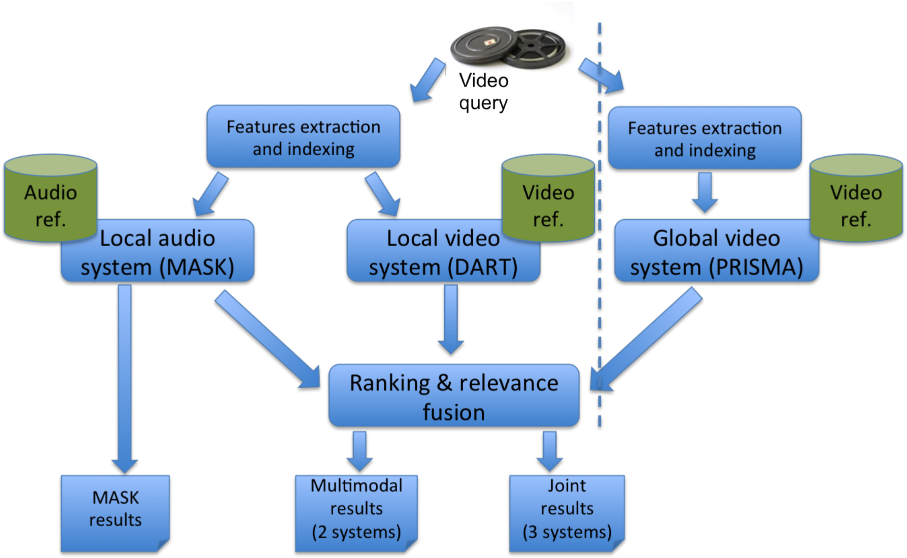


Fig. 1. *Blocks diagram of the system developed for Trecvid 2011, and the submitted outputs.*

be searched in them, considering that both audio and video transformations could have been applied to the videos, and without knowing where the copied segments are to be found within the query videos.

The system we contributed this year is an evolution of last years' system [2] . In addition, one of our submissions is the result of a collaboration with another participating team. Our system uses multimodal cues by fusing, at the decision level, the results of an audio and video systems. The video system is based on the Visual Search Engine (VSE) we used in 2010, although this year we have re-architectured it to eliminate any unnecessary intermediate steps resulting from the adaptation of our in-house image-based similarity system [3] to process full videos. In doing so, we have been able to eliminate many intermediate files and to make the system cleaner and more compact (all into one single software project). The audio system uses totally novel features called MASK (Masked Acoustic Spectral Keypoint features) which are local features extracted in the spectral domain of the video's audio track. The algorithm we use for matching the acoustic segments is based on the 2010 system, although with some modifications to account for the fact that MASK features are not extracted at constant time intervals, unlike the binary fingerprints, used last year.

Both modalities are executed independently and a list of possible reference video matches are obtained for each query, together with the matching segments and a score. Finally, a fusion algorithm we evolved from last year's system is used to intelligently merge both lists and obtain a resulting list with the fused results. Note that in all our submissions we have followed NIST's recommendation to produce multiple possible results for each query (in our case we produce 20 results for every query) in order to conveniently compute DET curves from the results. We are aware that this procedure harms our global performance given the much higher cost of false alarms versus the cost of missed matches.

In order to explore the possibilities of our fusion algorithm this year one of our contrastive submissions is based on the fusion of Telefonica's local audio and video outputs with PRISMA's global video outputs. As shown already in [4] for NIST-TRECVID 2010 results, the fusion algorithm is able to obtain better results when combining multiple modalities, given that they can bring orthogonal information not existent in the other modality outputs. Using Trecvid's 2011 data we found that the combination of 2 multimodal local feature approaches with a global feature approach can obtain very prominent improvements compared to any of the individual modalities alone, shown by the very good results we obtained with the "joint" system.

The remaining of this notebook paper is structured as follows: first we describe the overall system architecture and different monomodal systems we developed for the evaluation. Next we describe the fusion algorithm used to bring together all results into a single output file. Next we present our evaluation results and we perform an exhaustive evaluation of the fusion algorithm y using 17 system outputs of 10 participating teams. We finally we draw some conclusions from the evaluation and results, and talk about future work.

II. Telefonica Research Multimodal Video Copy Detection

The system presented by Telefonica in the video copy detection task at Trecvid is based on the fusion at decision level of several monomodal systems, as shown in Figure 1. Both the indexing and retrieval process is started by the extraction of features from the videos (reference or query) involved. This year we have used the same local video features as last year and have experimented with a totally novel local audio feature called MASK, described below. In addition, for this year's submission we have collaborated with Juan Manuel Barrios at PRISMA group in order to generate one of the contrastive submissions as the fusion of our monomodal systems with their global video-based system. Such collaboration has been encouraged by the good results we obtained in the fusion of Trecvid 2010 results shown in [4].

Once all features have been extracted they are either stored into the reference databases (different databases have been used for each modality, depending on the particular characteristics of each system) or used in the matching algorithms to find putative copies. In the later, each of the monomodal systems performs an independent search over their database, with different techniques, in order to obtain a list of $N_k = 20$ matching reference videos. For each possible match the systems return the start-end time of the matching segments both in the reference and query videos, and a score. Scores are not comparable among modalities, but are comparable within its modality.

The last step in the process is the fusion of all monomodal results. As explained in [4] and summarized below, the fusion algorithm takes into account the relative scores in each modality, the rank of each match within the list of matches in each modality and the overlap between matches across modalities to produce a final list of up to 20 matching segments. Although it could be thought that performing the fusion at the decision level has some shortcomings with performing a fusion at earlier stages, in our system this worked very well for several reasons: a) we can fuse different modalities, with no restriction on the number; b) no special care needs to be given to the way features in each modality have been extracted, or their matching scores; c) subsystems built for each modality can be implemented in very heterogeneous ways.

Next we will describe in some detail the three monomodal systems we used for this year's submissions. The Local audio and video systems together with the fusion algorithm were developed entirely by Telefonica Research, while the global video system was entirely developed by PRISMA group who then contributed the matching results for the fusion.

III. Local-Audio Copy Detection System

A. *MASK feature extraction*

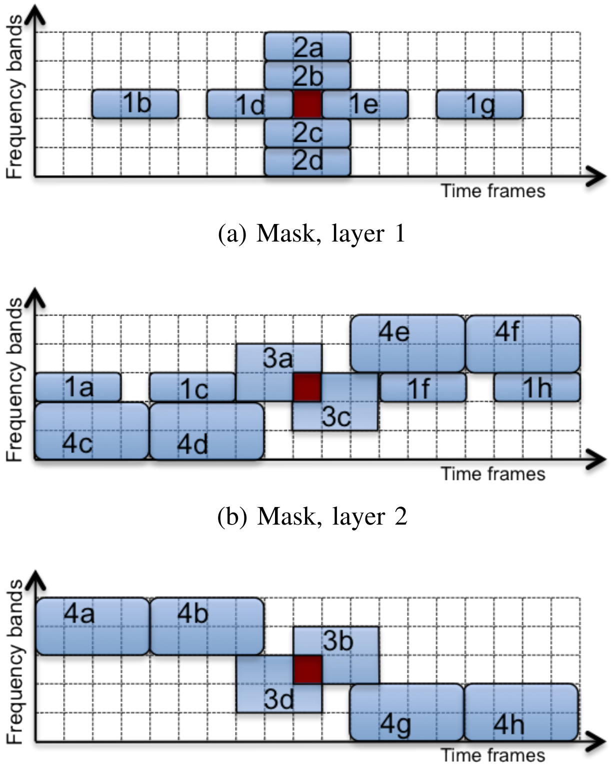
For this year's submission we have implemented a novel acoustic fingerprint that is localized in time and frequency and is capable of discriminating between different acoustic contents. The novel fingerprint is called MASK, which accounts for Masked Acoustic Spectral Keypoint Features. It derives from the observation that after a transformation has been applied to the audio signal usually the spectral peaks remain quite similar to those of the original signal. Such information is also exploited in some well-known fingerprints like the Shazam Fingerprint [5] or the Phillips fingerprint [6] although all these fingerprints differ in the way the information is encoded. The Shazam fingerprint encodes the position of some "anchor" peaks and their distance to other peaks around them. On the other hand, the Phillips fingerprint encodes the energy variation across different MEL bands, indirectly encoding both maxima and minima positions.

In MASK we tried to obtain an encoding that contained what we considered best of each of the previous and other implementations. To do so, we encode information around local maxima in the spectrogram, but we do not rely on its relationship to nearby peaks, as we consider that the resulting fingerprints would be less robust given that at any time any of these two peaks could shift or disappear. Instead, we encode the energy differences between regions around each peak. In particular, for any given signal we first compute the Fast Fourier Transform (for 100ms of audio, computed every 10ms) and apply the MEL-filterbank analysis to obtain a total of 32 bands (like in Phillips fingerprint). Next, we find the peaks in the MEL-spectrogram, i.e. the MEL band vs time index where the signal has a higher energy in comparison to all neighboring signals. In addition to this condition, we also applied other rules like a temporal masking region in order not to allow for two peaks to appear too close together.

Once the spectrogram peaks have been detected we apply a mask centered at each of the salient peaks. This defines regions of interest around each peak that are used for encoding the resulting binary fingerprint. A region in the mask is defined as either a single time-frequency value or a set of spectrogram values that are considered to contain similar characteristics (they are usually contiguous in time and/or frequency). The encoding is carried out by comparing the differences in average energies between certain region pairs. When a region is composed of several values, its energy is represented by the arithmetic average of all its values. The different regions defined in the mask are allowed to overlap with each other. The optimum location and size of each region in the mask, as well as the total number of regions, can vary depending on the kind of audio that is being analyzed and the number of total bits we desire for the fingerprint. The particular mask we used for TRECVID this year is shown in Figure 2. This mask covers 5 MEL frequency bands around the peak – 2 bands above and 2 bands below – and extends for 190ms – 90ms before and 90ms after. Different regions grouping together several spectral values are labeled using a numeric value followed by a letter.

Note that when a salient peak is found either at the band N-1 or at band 2 (i.e. with only one band above or below it) the mask in Figure 2 can not be placed correctly centered around that peak as either the first or last rows would fall outside of the spectrogram limits. In such case we duplicate the values of the first/last available band to cover the inexistent values for the first/last mask rows. We define the regions and the final fingerprint in a way that such redundancy do not affect much the properties of the resulting fingerprints.

Next, we construct the fingerprint characterizing each peak by combining both the index of the frequency band where the peak being described was found and the information from the masked area around it. In our work we aim at the construction of an up to 32 bits long fingerprint, which is sufficient for the indexing and retrieval of a very large number of audio documents. Future extensions to 64 bits are possible and very straightforward by just redefining the mask and extending the set of comparisons between its regions. The information of the masked area around each fingerprint is encoded by comparing (in pairs) different regions pairs from the masks above. For every pair we set a particular bit to 1 if the first region has greater average energy, or to 0 otherwise.



(a) Mask, layer 1

(b) Mask, layer 2

(c) Mask, layer 3

Fig. 2. TRECVID mask covering a region of 5 bands per 19 temporal frames, split into 3 layers to better observe the overlapping regions

B. Acoustic matching algorithm

Like in last year's evaluation, comparison between reference and query is performed in a one-to-one basis (future work still includes the indexing of all reference fingerprints for faster search). First we index the fingerprints corresponding to the reference into a hash table, storing also the time frames where they occur. Next we use an algorithm inspired on the temporal matching algorithm proposed in [7] in order to group the matches into segments. Given the list of matching MASK reference keypoints for every query keypoint, the algorithm uses the difference between the time-stamps of each query-reference keypoint pair to construct a histogram of matching time-differences. Assuming a perfectly aligned matching between query and reference segments (i.e. there is no time warping or it is negligible) we will observe a big peak in the time-difference that optimally aligns query and reference. In top of this algorithm we apply two variations. On the one hand, we explicitly consider a plausible small jitter/warping of the alignment by considering for each time-difference not only its exact value, but also the values around it. On the other hand, for a given time-difference we we consider two adjacent matching keypoints (either in the reference or in the query) belong to the same matching segment only if their time distance is below a maximum non-matching time (which we set to 5 seconds). Any bigger distances prompt the algorithm to create a new matching segment in the same histogram position.

Once all keypoint pairs have been inserted into the histogram we obtain a matching segment's start-end position and a score. Like last year, we are not very confident that this score can optimally represent the distance between query and reference segments as it contains only the evidence of the query keypoints that match exactly the reference keypoints. In the case of a clean audio this evidence could be enough, but in the case of strong transformations (like some overlapping noises in TRECVID) there will be many non-exact matches to be considered. For this reason we perform a post-processing step to compute a more accurate score. Given that the MASK fingerprints are not extracted at regular time intervals we need to introduce a new layer of complexity in finding the optimum alignment of non-exactly-matching keypoints when comparing the query and reference matching segments. To do so, we first align all keypoints from the 2 matching segments at the optimum time-difference obtained in the previous step. Then, for every keypoint in the query that does not have a corresponding exact match we find the reference keypoint with smallest Hamming distance within a small window around the position of the keypoint. At the end of the process we have a Hamming distance for every query keypoint that we can use to estimate with greater detail the density of matching points in the matching segment.

IV. Local-Video Copy Detection System

This year's local video system is based on last year's submission with several engineering changes to convert the system from image to video processing and to avoid processing errors when lots of data flows through data networks. Next we review the system's main characteristics. For a more detailed description please refer to last year's submission [2].

A. DART Features Extraction

First an FFMPEG-based module extracts the video track from the input video (both the reference and the query videos) and extracts one keyframe per second into memory. We chose not to deal with any shot-boundary detection system in order to avoid errors in this step and ensure completeness of the information extracted from the videos. As a downturn, our database becomes very big and we need to split the TRECVID collection into chunks in order to fit the data into memory on commodity machines.

In each keyframe we perform a detection of inserted static text and patterns, which are an important source of errors for any local feature we tested. The detector we used operates by sliding a temporal window of a few keyframes along the video. For every keyframe an initial mask corresponding to static regions is created by finding pixels whose intensity has zero standard deviation within the temporal window surrounding the current keyframe. Then, a dilation operator is applied to the initial mask in order to ensure appropriate margins surrounding the static patterns and also to fill out possible inner holes. Conveniently, the method also masks regions close to the black layout borders that are not very useful for matching.

In addition, many of the TRECVID videos contain moving text and subtitles. Like with the fixed text and logos, local features extracted from these added patterns are prone to match very well with similar patterns inserted in totally different videos, therefore raising the number of detected false alarms. In order to avoid extracting features in the areas with subtitles and moving texts we can not apply the same technique used with static patterns as they change many times along the duration of a video or they are sometimes semitransparent. Instead, we have developed a very simple yet effective dedicated subtitle region detector that relies on the analysis of the spatial density of vertical edges within every single keyframe.

In this method, first vertical edges (low-to-high or high-to-low transitions within every image row) are detected using the Sobel operator and binarized with respect to a predefined threshold. A pixel is classified as part of a textual region if the density of the edges within a sliding window centered at the pixel of interest is higher than a predefined threshold. Once all pixels are classified as text/no-text, the resulting initial mask is extended by applying the dilation operator within every row to ensure a secure margin around textual regions and also to fill out holes between or within letters. Since the above method relies only on the presence of vertical edges it works well for solid and transparent letters. Like in last year's system, we used different parameters for the top part of the keyframe than in the bottom part, where subtitles have a higher prior probabilty of appearance.

Finally, we extract our local features in the regions that have not been masked in the previous processing. We use Telefonica's DART features [8]. In our experiments with image databases typically DART performs better or comparable to Scale Invariant Feature Transform (SIFT) [9] and Speeded Up Robust Features (SURF) [10] in terms of repeatability, and precision vs recall [8]. Moreover, it is very attractive in the context of the video copy detection task because of its very low computational cost (6x faster that SIFT and 3x faster than SURF), and compactness (only 68 components).

For TRECVID keyframes we set the maximum number of extracted keypoints to 400 in reference videos and 800 in query videos. Given all available keypoints in a keyframe we rank them according to two factors, and select the most representative according to the maximum numbers indicated above. The first factor we take into account is the scale at which the keypoints have been extracted. We consider keypoints obtained at higher scales to be more resilient and interesting. The second factor we apply is a temporal keypoint consistenty across adjacent keyframes, considering as most relevant those keypoints that have similar keypoints at the same locations in one of both neighboring keyframes.

Differently from last year's system, all the previous steps are now done inside a single software, allowing to obtain a single binary feature file for every input video. Given that the system we use was adapted from a more generic image matching system, we initially had to store tons of temporary files in disk corresponding to each keyframe and derived binary keypoints, that lead to many problems in storage and management of execution and network errors.

B. Visual Matching Algorithm

Once reference and query keypoints have been extracted we run the matching algorithm to find possible matching segments of a given query in the reference database. Generally, to enhance the matching speed of the system we store the reference keypoints in memory before a query is searched through them. In the matching process we are using commodity machines with less than 2GB of RAM memory, thus we need to split the reference database into chunks and perform a query search over different processing nodes. The result of each matching node is a list of potential reference matches that we then join together, rerank and trim to $N_k = 20$ final results.

V. Global-video Copy Detection System

The visual global features module was developed by PRISMA Research Group at the University of Chile [11]. It divides the detection process in five tasks: Preprocessing (which minimizes the effect of visual transformations), Video Segmentation (which partitions every video into segments), Feature Extraction (which represents each segment with one or more descriptors), 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 the location and score for each copy candidate).

For this submission each video was divided into segments of 333 ms length. Two visual global descriptors were extracted for each frame and averaged for each segment: `Ehd` which divided every frame into 4×4 blocks and for each block measured the distribution of 10 orientation of edges; and `Rgb` which divided every frame into 4×4 blocks and for each block calculates a 4-bins histogram for each of the Red, Green and Blue channels.

The distance between video segments is defined as a weighted combination of global descriptors. The weights were automatically calculated using the Weighting by Max-τ algorithm with α=0.001. The approximate search retrieved the k=10 nearest neighbors using an approximation parameter T=1% with 5 pivots. Finally, the copy localization selected the 20 copy candidates with higher score for each visual query video (all these algorithms are detailed in [11]). This list of the best candidates were the output of this system to the fusion module.

VI. Multimodal Fusion Algorithm

Every monomodal system described above takes a decision on which reference video segments optimally match the query video, together with a matching score. In this final module we perform the fusion of all these results and obtain a final multimodal list of resulting matches. The algorithm reviewed here is similar to the one used in TRECVID 2010, where we have slightly modified to make it robust to a) queries from a single modality with fewer results than normal; and b) queries with no results at all.

Generally, once all individual monomodal systems have finished, they output a result consisting of the score-ranked N_k-best reference matches. By fusing all these results into a single output we are able 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. As a side-product of the algorithm implementation, all resulting 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 or their initial scores. In our system we decided to set both N_k and N to be 20. We are fully aware that making $N = 20$ we are more prone to false alarms jeopardizing our final NDCR results (mostly in the NoFa case) but we believe it is more usable to have 20 ranked results for many different applications.

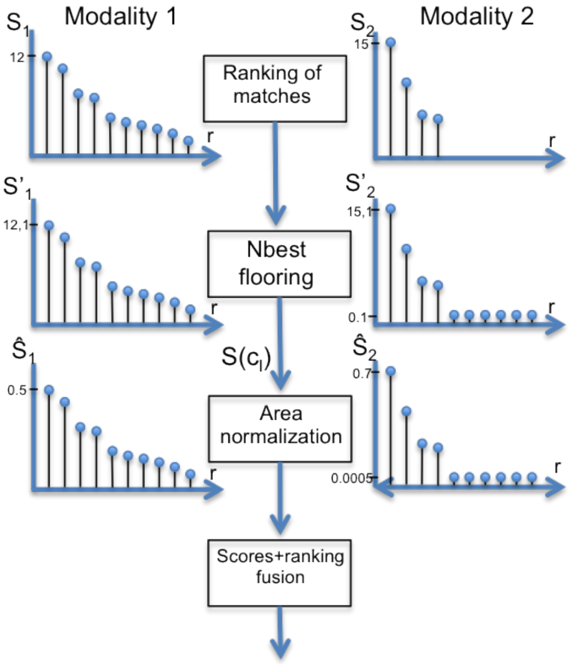


Fig. 3. *Steps involved in the multimodal fusion algorithm.*

Figure 3 shows the main blocks that can form the proposed multimodal fusion algorithm. The input of the algorithm can be any number of individual system outputs, although in Figure 3 we just show two for convenience and space limitations. Next we describe in detail each of the steps involved in the fusion.

A. 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 and the distribution of scores for every modality will usually be different. In order to avoid problems in the subsequent steps we perform a simple scores normalization dividing each score by the median score of the scores for all queries in a given modality. Although we could do much more complicated normalizations (e.g. normalizing the scores distributions) we found the median score to be a good trade off between simplicity and correctness of results.

In the next step we introduce a flooring factor α. In the case that any of the input modalities was 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 forcing all modalities to have N results by extending the number of results to this number, with α score. This flooring has two functionalities: on the first hand it acts as a normalization threshold for the scores when applying the normalization step that follows. On the other hand it is a simple way to deal with those modalities that do not provide any result, i.e. their N-best results are all at score α, thus penalizing the final score of any other results from other modalities.

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. For example, if one or a few scores show much higher values than the rest in a particular modality, they will retain such difference once normalized and will remain high when compared with other modalities. On the contrary, when all N_k scores have very similar values, the normalized scores will be close to $\frac{1}{N_k}$ and will not stand out across modalities. Note also that by using the α flooring we ensure that modalities with very few (sometimes only one) very low scores do not turn to be very prominent in the fusion as they get normalized with the accompanying α values.

B. 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

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

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

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

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, $\hat{S}_k[1]$, which is equal to α in modalities with no results for that query, 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.

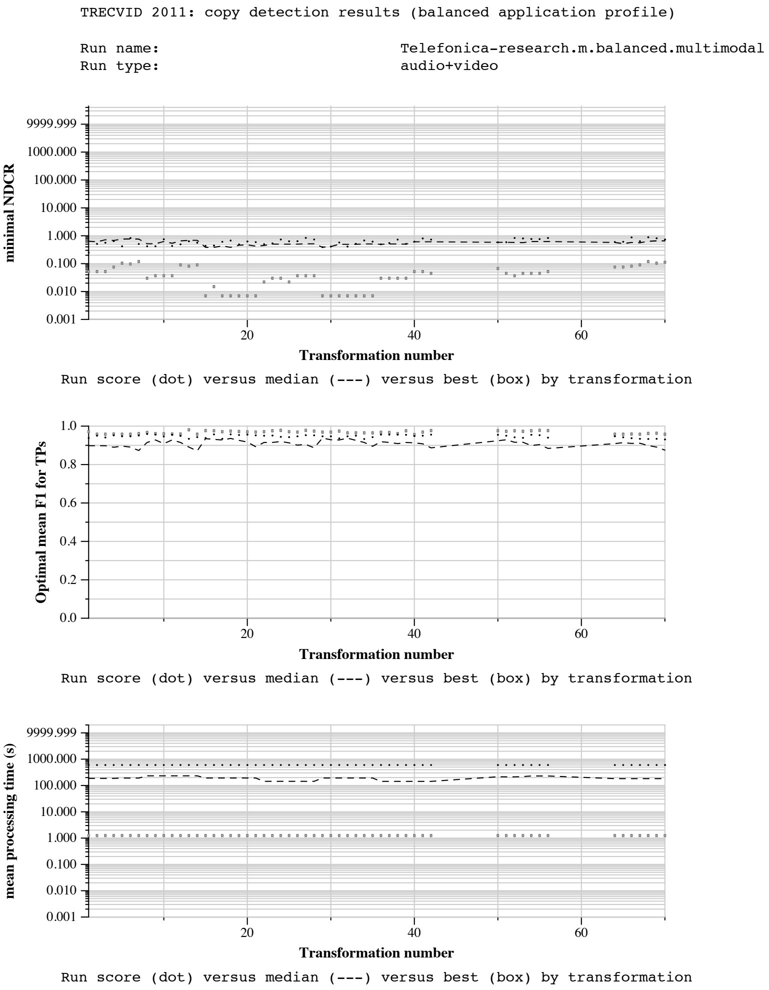
VII. Evaluation results

In this section we present the official results we obtained running the 4 submissions to TRECVID. Figures 4 and 5 show the results for the primary submission, which fuses the results of our audio and video engines, for the balanced and no false alarm profiles, respectively. Figure 6 shows results of the audio-only system using MASK features and the balanced profile. Finally, Figure 7 shows results of the joint submission with PRISMA group, which fuses the audio and video systems from Telefonica with the video-only system from PRISMA. The metrics shown in the plots correspond to the standard metrics used by NIST. These are the NDCR, the F1 and the running time. The NDCR corresponds to a weighted sum of the cost of misses and false alarms in the detection of copies. Both in the Balanced and in the no false alarms profiles the cost of false alarms is much higher than the cost of misses. In our system we are not interested in the no false alarms profile as it does not conform to any real-live task we are interested in applying these algorithms to. For this reason we are not taking any special measures to limit the false alarms in the system, except from changing the detection threshold. Furthermore, for all systems we return the 20-best matching reference videos, even though NIST specifically states that only one video copy might be found for each query. This really jeopardizes our nofa results, which leads us to focus on the balanced results (we submitted a nofa result as it was a requirement for participation in the evaluation).

Table I summarizes the average optimum results for our submissions. We can see how the *multimodal NoFa* submission performed quite poorly in terms of NDCR given the high penalties resulting from false alarms. On the contrary, the *multimodal* system performed quite well in F1 both in the *NoFa* and the *Balanced* profiles (nearly identical results). We can see that the F1 is even higher for the *joint* submission, resulting in one of the highest F1 results this year, but it decreases in the audio-only (*MASK*) submission. We can see how both the NDCR and the F1 show improved results as more modalities are fused, as *MASK* has the worse results with only one modality involved, while joint has the best results, with 3 modalities. This indicates that the fusion algorithm is working well in combining the non-correlated information that is brought by the different systems involved.

TABLE I
Summary of TRECVID Video Copy Detection Optimum Results

System	Profile	Min NDCR	Opt. F1
Multimodal	NoFa	57.768	0.948
Multimodal	Balanced	0.610	0.947
MASK	Balanced	0.662	0.729
joint	Balanced	0.268	**0.957**



```
TRECVID 2011: copy detection results (balanced application profile)

Run name:                          Telefonica-research.m.balanced.multimodal
Run type:                          audio+video
```

Run score (dot) versus median (---) versus best (box) by transformation

Run score (dot) versus median (---) versus best (box) by transformation

Run score (dot) versus median (---) versus best (box) by transformation

Fig. 4. *Results for primary (multimodal audio+video) submission, balanced profile, optimum results.*

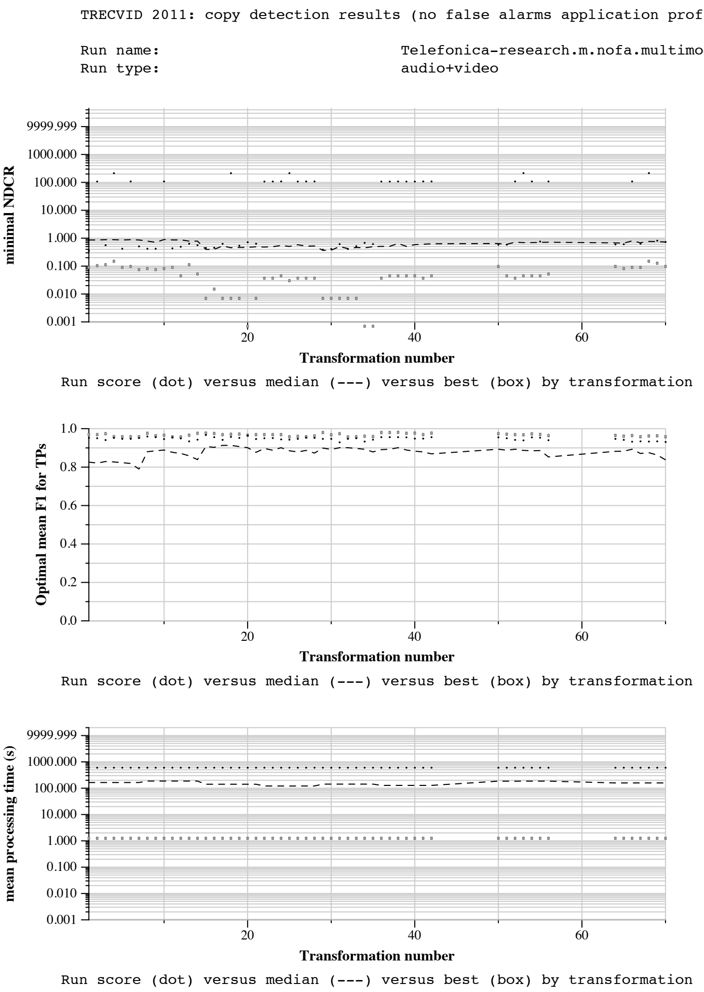


Fig. 5. *Results for primary (multimodal audio+video) submission, nofa profile, optimum results.*

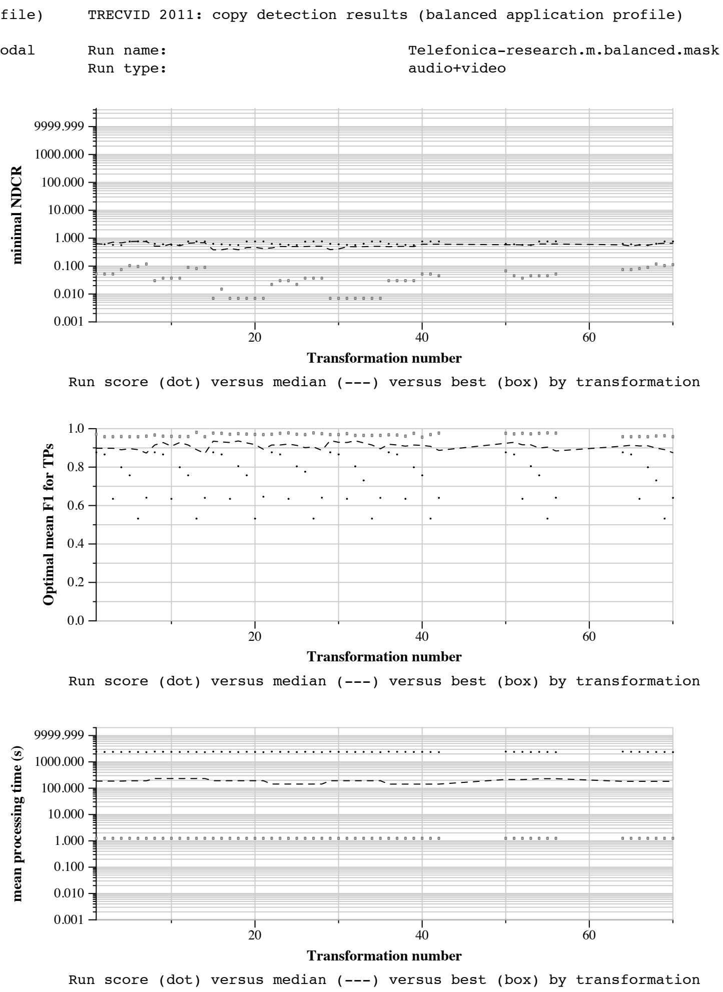


Fig. 6. *Results for contrastive audio-only submission, balanced profile, optimum results.*

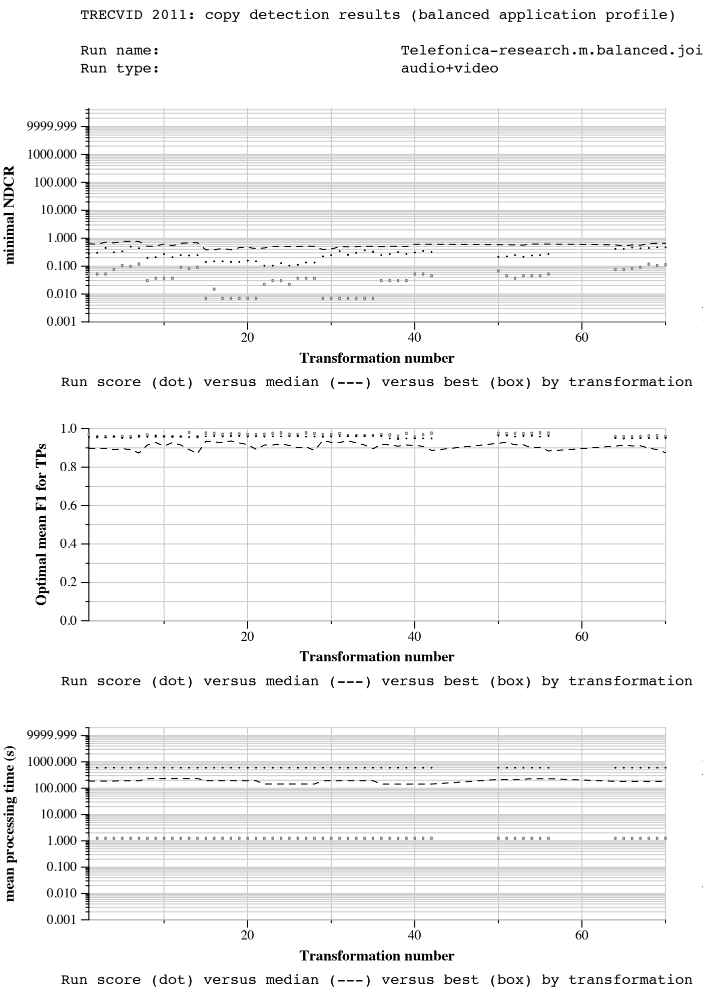
VIII. Overall Fusion Experiments

As explained in above, the fusion algorithm we used in this year's evaluation is able to effectively combine the results from different individual systems and obtain an enhanced fused result. In our experiments we used it with only 3 different systems, for this reason after results were returned to participants we requested participating teams to share with us their submissions in order to try the fusion algorithm is a larger scale. A total of 10 teams shared their submissions from which we selected 17 runs belonging to the balanced profile submissions. Individual evaluation results obtained by these systems range from 0.053 and 0.99 Optimum Balanced NDCR, as shown in Figure 8. To preserve the identity of the submitting participants we label the system outputs from 1 to 17.

Note that the described fusion algorithm takes advantage of the rank of the different matching segments for a given query. We noticed that several systems consistently returned a single (or very few) results per query. While forcing the output of a system to return a single result (or no result at all) helps reduce false alarms, it is not optimal for the fusion algorithm, which is not performing to its best in these cases.

Figure 9 shows the results of the first experiment, where we plot the optimum NDCR scores obtained by the fusion between 2 and all 17 system outputs. The order followed to perform the fusion was in order of individual NDCR scores, from best to worst. We see how the fusion normally improves the overall NDCR score except for systems 5 and 15.

Next, Figure 10 shows the relative importance of each submission in the overall fusion by showing the result of fusing all systems except one. By comparing the results with the overall fusion NDCR obtained in the first experiment (0.0333) we can see that some system outputs were able to contribute



```
TRECVID 2011: copy detection results (balanced application profile)

Run name:                              Telefonica-research.m.balanced.joint
Run type:                              audio+video
```

```
Run score (dot) versus median (---) versus best (box) by transformation
```

```
Run score (dot) versus median (---) versus best (box) by transformation
```

```
Run score (dot) versus median (---) versus best (box) by transformation
```

Fig. 7. *Results for contrastive joint Telefonica-PRISMA submission, balanced profile, optimum results.*

to the overall fusion (leading to a higher NDCR once they are taken out), while some harm the system (by leading to a better NDCR value once take out). The clearest example of the second case is system 15 (also seen in the previous experiment), without which we obtain an NDCR score of 0.0206.

Finally, figure 11 shows results for a final experiment where we start from the fusion of all systems and finish with only 2. Unlike in the first experiment, in here we draw the order of exclusion from the fusion by first eliminating those systems that in the second experiment showed to contribute less to the fusion (i.e. starting at system 15), an finishing with those two that contributed the most (i.e. the NDCR scores deteriorated the most when taken out).

Results show that the best score we obtain is once we eliminate the third system, with an NDCR score of 0.0195. Interestingly, NDCR scores degrade steadily until only 5 systems are used, and when fusing 6 systems we already obtain better results (0.046) that the best system output we used.

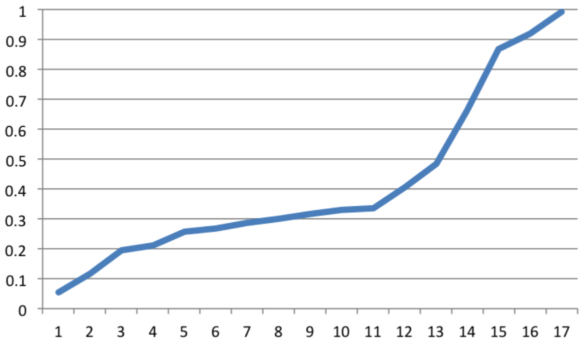


Fig. 8. *Individual system results used in the fusion experiment.*

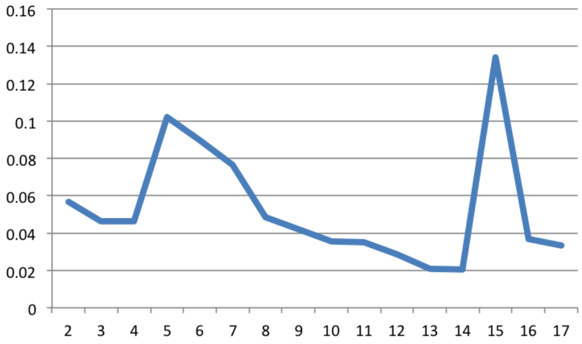


Fig. 9. *Overall fusion experiment iteratively including systems to the fusion.*

IX. Conclusions and future work

This year through the participation to the NIST-TRECVID evaluation we have focused on the following: 1) we have reworked the video-based system to deal with the videos more effectively; 2) we have investigated a novel local-audio feature called MASK and used in for the TRECVID task; and 3) we have improved the multimodal fusion algorithm we presented last year and submitted results of our systems jointly with the results of the PRISMA group. Over all we are pleased to see that our submissions obtained very good results both in NDCR and F1 for the BALANCED task, which is the one we see as closer to real-life applications. We are also pleased to see the system slowly becoming more mature, as now it is based solely on proprietary technology, and is becoming a robust software entity. Next steps we are working on is the scalability to amounts of multimedia content far beyond those used in TRECVID in order to apply these systems to larger applications.

X. Acknowledgements

We would like to acknowledge those teams sharing their submission outputs with us for performing the fusion algorithm analysis shown in this paper.

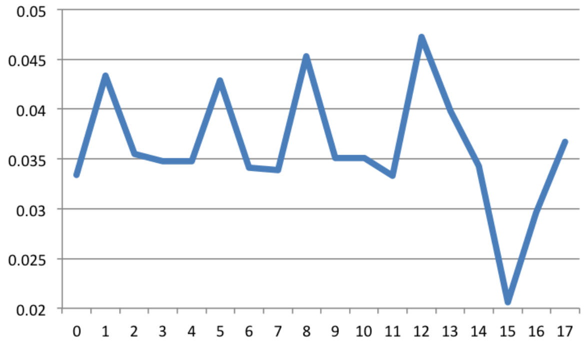


Fig. 10. *Overall fusion experiment excluding from fusion one system at a time.*

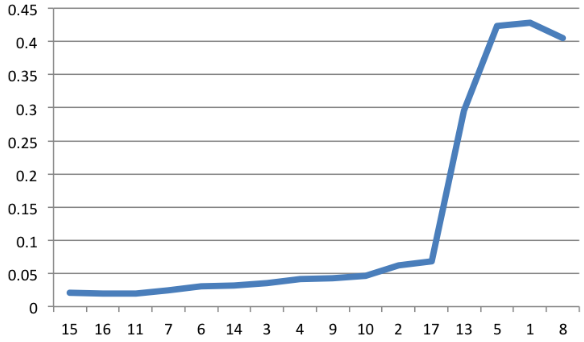


Fig. 11. *Overall fusion experiment iteratively excluding those systems with worse contribution to fusion.*

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

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

[3] T. Adamek and D. Marimon, "Large-scale visual search based on voting in reduced pose space with application to mobile search and video collections," in *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, july 2011, pp. 1 –4.

[4] X. Anguera, J. M. Barrios, T. Adamek, and N. Oliver, "Multimodal fusion for video copy detection," in *Proc. ACM Multimedia*, 2011.

[5] A. Wang, "An industrial strength audio search algorithm," in *Proc. ISMIR, Baltimore, USA*, 2003.

[6] A. K. Jaap Haitsma, "A highly robust audio fingerprinting system," in *Proc. International Symposium on Music Information Retrieval (ISMIR)*, 2002.

[7] Z. Liu, T. Liu, and B. Shahraray. (2009, November) http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/att.pdf. Trecvid 2009 Online proceedings. [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv9.papers/att.pdf

[8] D. Marimon, A. Bonnin, T. Adamek, and R. Gimeno, "DARTs: Efficient scale-space extraction of daisy keypoints," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. European Conference on Computer Vision (ECCV)*, May 2006. [Online]. Available: http://www.vision.ee.ethz.ch/~surf/index.html

[11] J. M. Barrios and B. Bustos, "Competitive content-based video copy detection using global descriptors," *Multimedia Tools and Applications*, pp. 1–36, 2011.