Automatic Textual Annotation Of Video News Based on Semantic Visual Object Extraction

Nozha Boujemaa, Francois Fleuret, Valerie Gouet, Hichem Sahbi
INRIA - IMEDIA Research Group - Domaine de Voluceau - BP 105 - F78153 Le Chesnay Cedex

ABSTRACT

In this paper, we present our work for automatic generation of textual metadata based on visual content analysis of video news. We present two methods for semantic object detection and recognition from a cross modal image-text thesaurus. These thesaurus represent a supervised association between models and semantic labels. This paper is concerned with two semantic objects: faces and Tv logos. In the first part, we present our work for efficient face detection and recogniton with automatic name generation. This method allows us also to suggest the textual annotation of shots close-up estimation. On the other hand, we were interested to automatically detect and recognize different Tv logos present on incoming different news from different Tv Channels. This work was done jointly with the French Tv Channel TF1 within the "MediaWorks" project that consists on an hybrid text-image indexing and retrieval plateform for video news.

Keywords: Visual content analysis, semantic feature extraction, automatic textual metadata generation, text-image thesaurus, application to video news.

1. INTRODUCTION

In this paper, we address the detection and the analysis of two major semantic object in the video news archives: faces and Tv chanel logos. For both logos and faces, we consider a thesaurus identification approach to automatically suggest textual annotation from this object recognition process. We notice that we did not use any image caption analysis or audio transciption for person names as done in former approches such as Name-it [5]. Moreover, we can deduce textual annotation of clos-up from face detection parameters. For logos identifiction, local image descriptors was invetigated for detection and matching process. Performance and limits of this approach are presented.

2. AUTOMATIC TEXTUAL ANNOTATION FROM FACE ANALYSIS

Automatic extraction of information related to the presence of individuals in images is of major interest for indexing and searching video databases. The work presented in this section has served for face's information annotation in the TF1 (the French TV channel) video news archives within Mediaworks project. This task has been adressed by coupling a rapid face-detector with a novel face recognition algorithm presented in the following sections which allows us to generate a textual close-up estimation and to label faces by their identity. This is not an easy task since faces are considered as semi-rigid objects subject to variations in rotation, scale, lighting conditions, non-linear deformations for instance expression, and occlusion.

2.1. Face detection with textual close-up estimation

Recent advances in statistical learning led to the design of efficient face detectors, both in term of error rates and complexity. We define here face detection as an automatic process able to locate precisely all frontal-viewed faces in an image. The standard way to address this challenge consists in building a two-class classifier which labels patches of scenes as either "face" or "non-face". This classifier is then applied everywhere in the scene, at various scales to cope with faces of various sizes. The main challenge in the design of this detector is to obtain a very low false-negative error rate (i.e. very few missed faces), even if the natural prior is highly unbalanced and the "face event" very rare. From a pragmatic point of view, if one was sampling thousands of scene patches randomly on the web, he would likely not catch any face.

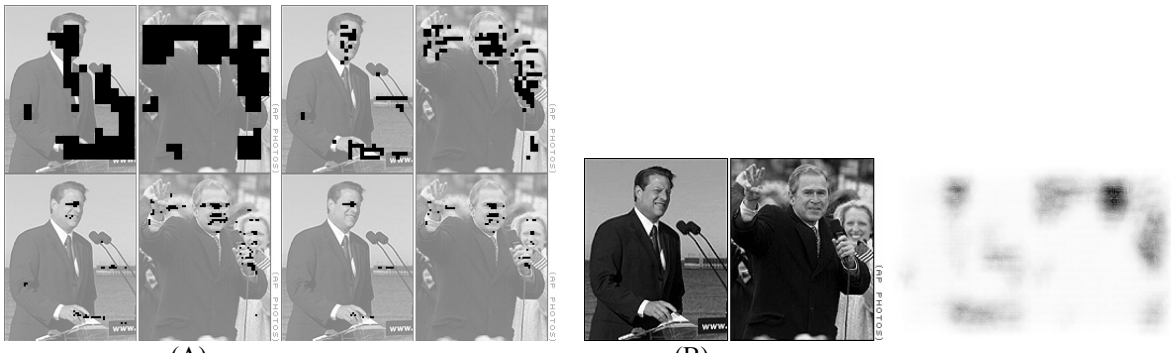
Our approach is based on the generic idea of coarse-to-fine strategies. It consists in a hierarchical structure both to model the signal we are looking for (in that case faces) and to process the data. The algorithm we have developed has the form of a hierarchy of simple classifiers, each dedicated to population of face pictures more or less constrained in pose. Precisely, each classifier is built from a training set generated by forcing the locations of the eyes of all the face examples to live in a certain domain. This domain is very large for the top-classifiers of the hierarchy (which means that the distance between the eyes, the tile and even the location of the faces they are supposed to spot can vary a lot), while it is highly constrained for the bottom classifiers. Thus, the last classifiers of the hierarchy have to respond only if a face whose eyes are at precise locations is present (Figure 1.A).

The first classifiers of the hierarchy have a very high false-positive error rate, which means that they are easily fooled by face-like structures appearing in the scene. They reject nevertheless huge chunks of the image with a very low processing cost. Parts of the scene not rejected by those first classifiers are checked by the classifiers in the levels below in the hierarchy. The deeper a classifier, the more dedicated it is. Thus, classifiers in the deep levels can look for structures (eyebrows, mouth, nose, etc.) at exact locations, and can achieve very low false-positive error rates. In term of complexity, each level of the hierarchy has two times more classifiers than the preceding layer to address all the possible face shapes, which leads to a complexity exponential with the depth.

This technique reduces dramatically the processing time, as it concentrates the computation on ambiguous areas of the scene. Uniform areas (sky, walls, clothes) are rejected early in the process, while less common but complex parts (vegetation, patterns, texts) have to be processed intensively (cf. Figure 1.B). This behavior is very similar to our own cognitive process. Consider the difference in the way you look for faces in an empty scene (for instance in the desert) or in a complex one (jungle). You will notice that while in the first case you directly jump to the faces, you need to scan carefully the complete scene in the second.

The detection process provides a list of alarms in the scene and each alarm is characterized by the sequence of classifiers which have responded positively. Each classifier is dedicated to a certain family of faces, and implicitly to a family of face pose. Finally, one can associate to every alarm a list of average face poses, defined as the eyes locations in the image plan. This information is obtained for no additional cost. A post-processing consists in doing a bottom-up clusterization of this set of alarms to group similar ones in term of pose into clusters. The final result is the list of centroids of clusters of sufficient cardinality. This gets rid of the few remaining false alarms and provides accurate and reliable pose estimations for the detected faces.

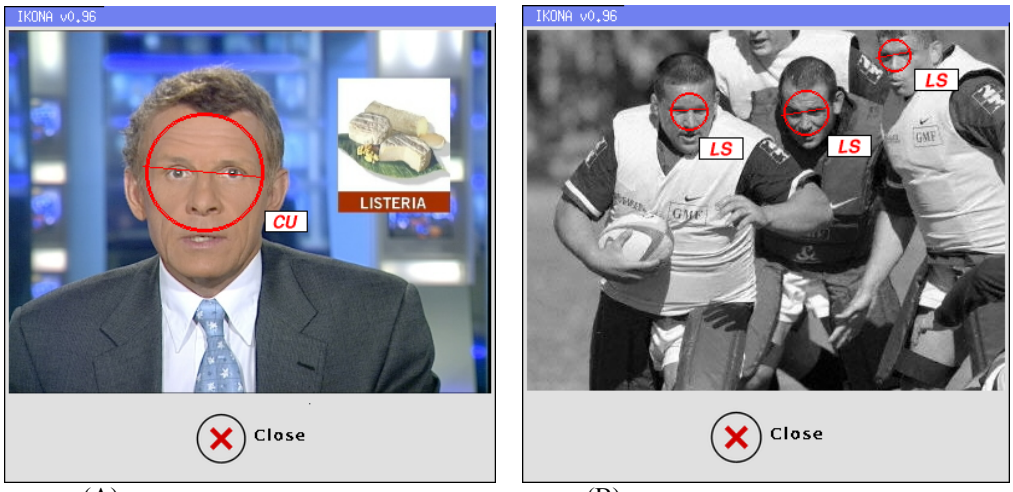
Also, we get an estimation of the size of each face defined as the distance between its eyes. Thus, beside the information about the location, we also get an automatic labeling of the type of shots. Depending on the distance between the eyes, relatively to the scene size, each face is qualified as Long Shot, Medium Shot, or Close Up (cf. Figure 2). From that information, we can provide the user with new requests based on the number of visible faces, their locations in the scene, and the type of shot.



(A) (B)

Figure 1. (A): Those pictures show what areas of the picture remain candidates for the locations of centers of faces after each level of processing in the hierarchy. The first level eliminates uniform areas but still generates false alarms as

soon as edges are present. When the process goes on, more and more alarms are rejected. **(B):** Those two pictures show the intensity of the computation. During the detection in the scene on the left, we count how many times each pixel is accessed. The picture on the right shows those numbers of accesses as gray levels. Darker parts correspond to large numbers of accesses while white ones correspond to very few numbers of accesses. As it can be seen, the computation accumulates on complex parts of the picture.



(A) (B)

Figure 2. (A): The face detection estimates automatically the precise eye locations. Besides providing information about the locations, it can also label each face as "Long Shot" (LS) if its relative size is small, "Medium Shot" (MS) if it is average and "Close Up" (CU) if it is big.

2.1.1. Face recognition and semantic interpretation

Existing methods for face recognition can either be classified as local or global. Local methods [1] [2] are based on extracting some face characteristics (points of interests or face components) which are used in a convenient way to estimate matching distances or some anthropometric characteristics of faces. Global methods map each face image into a space referred to as the feature space and use a similarity measure in order to retrieve and to label face images. Both global and local methods suffer from several drawbacks: global methods are very sensitive to the non-alignment of the facial components and need a careful pre-processing in order to register faces. On the other hand local methods require an accurate and automatic extraction of face components (eyes, nose, etc.) and this is known to be a difficult and an unsolved problem.

In this section, we propose a method for face recognition and labelling which is based on dynamic programming. This approach is robust to both photometric variations and non-linear face deformations since:

- We pre-process each face image in the database in order to estimate the saliency of each pixel belonging to this image and this is an invariant measure with respect to variations in photometry. The image which contains the saliency of each pixel is referred to as the **entropy map.**
- We extract systematically many sub windows from the entropy map and we use them in a matching process based on dynamic programming.

2.1.2. The entropy map

This is an invariant measure which provides for each pixel in a face image its saliency, i.e., its importance in the recognition process. For each pixel in a face image, we compute a local gray level histogram around it (see. Figure 2, left). Then the entropy of this histogram is estimated and the saliency of the underlying pixel is set to the value of the estimated entropy. Local regions with rich information such as the eyes, mouth and nose have an extended gray level

histogram and the underlying entropy value (and the saliency) is important whereas uniform local regions such as the forehead have a peaked histogram and the underlying entropy value is small.

Figure 3. Left: Different regions and their underlying gray-level histograms. Right: The entropy map computed on different face images.

The estimation of this map can be performed efficiently using an incremental updating of the local gray level histograms by subtracting the gray level frequencies of the removed pixel and adding the new gray-levels. The robustness of the entropy measure to linear photometric variations has been demonstrated in [3]. Figure 3 (right) shows faces with different skin and lighting conditions and their underlying entropy maps.

2.1.3. The matching process

Given a query image Q and a candidate image C from a corpus of face images, the matching process is based on extracting systematically sub-windows, from both the entropy maps of Q and C, which are used in the matching process. Let $X_1,\ldots,X_m$ (resp. $Y_1,\ldots,Y_n$) be the sub-windows extracted from the query (resp. the candidate) entropy maps. The matching process consists in finding (if possible) for each sub-window X_i from Q its underlying Y_j in C. The matching algorithm used is dynamic programming [4]. It is known that dynamic programming uses an ordering assumption, which states that: if X_i is matched with Y_j then X_{i+1} can be matched only with Y_{j+k} ($k > 0$). In practice, we can use this assumption for matching the extracted sub-windows since the pose parameters, resulting from the detection process, can be used to normalize faces and to make them upright.

The fact that only a subset of sub-windows from an image Q will match a subset of those of C makes it possible to overcome the occlusion effects due for instance to the presence of sun-glasses, scarves, etc. Furthermore, the matching process is robust against variations in expression since any local displacement of a face component (mouth, eyes) can be handled by the pseudo distance of dynamic programming, which is defined, between a set of two sub-windows $Q=\{X_1,\ldots,X_m\}$ and $C=\{Y_1,\ldots,Y_n\}$, as:

$$Trans(Q,C)=\sum_{i=1}^{m}\sum_{j=1}^{n}S_{ij}\ Sub(X_i,Y_j)+\sum_{i=1}^{m}I_i\ Ins(X_i)+\sum_{j=1}^{n}D_j\ Del(Y_j)$$

Here *Sub*, *Ins* and *Del* are respectively the substitution, the insertion and deletion metrics. The substitution metric tells us that two sub-windows (for instance two eye regions X_i and Y_j belonging to Q and C) are matched, so $S_{ij}=1$, whereas the insertion or the deletion metrics inform us about the presence of a sub-windows X_i (resp. Y_j) in the image Q (resp. C) which cannot be matched, so $I_i=1$ (resp. $D_j=1$). For example, this could be a mouth with a strange expression, or closed eyes, etc.

2.1.4. The annotation process

Each detected face image is processed by normalizing its pose and making it upright. Then, we estimate its entropy map, and we compute its matching pseudo distances **ì Transî** to the 300 faces belonging to the corpus. Given a face image Q from the video set and a face image C_{ij} (C_{ij} is the jth face image of the ith person in the corpus), the identity of Q is inferred by:

$$Identity(Q)=\arg\min_i\ \frac{1}{N_i}\sum_{j}^{N_i}Trans(Q,C_{ij})$$

Here N_i is the number of face images in the corpus which belong to the i^{th} person.

2.1.5. Results

Experiments have been conducted on the French TV channel (TF1) video set where we have **a corpus,** which contains 300 well-extracted face images corresponding to 15 persons from the French government. These images are used in order to find the identity of faces in the video set. More generally, this thesaurus approach is well suited for VIP recognition. This allows a central and updated people (textual and visual) reference for all archivists of TF1. The textual information can be more consistent than only names but also more complete information on function. The video set consists in a news stream of 50 minutes, which was broadcasted by TF1 on May 5th 2002. We sampled the video at one frame each 4 (s), resulting into 750 images containing 1077 faces and we run our face detector on the extracted frames. Some of these frames are shown in Figure 4. Figure 5 shows some results and the identity of different persons in the TF1 video set.

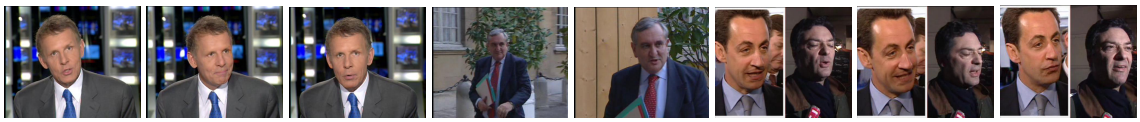


Figure 4: Some frames from the TF1 video set.

Figure 5: Some detection and annotation results. ì INCOî means that a region has not been recognized as a face and is annotated as background.

3. TV LOGOS DETECTION, IDENTIFICATION AND IMAGE ANNOTATION

Content-based image description can provide automatic textual annotation from other visual information, such as logos. In this work, we propose in this part an approach that allows automatic detection and identification of logos possibly contained in an image (for instance key frames of videos). The application considered here is the automatic characterization of TV video streams copyrights for the audio-visual domain. This approach requires the availability of a thesaurus of logos annotated with textual information (labels). Its principle is illustrated in Figure 6 below.

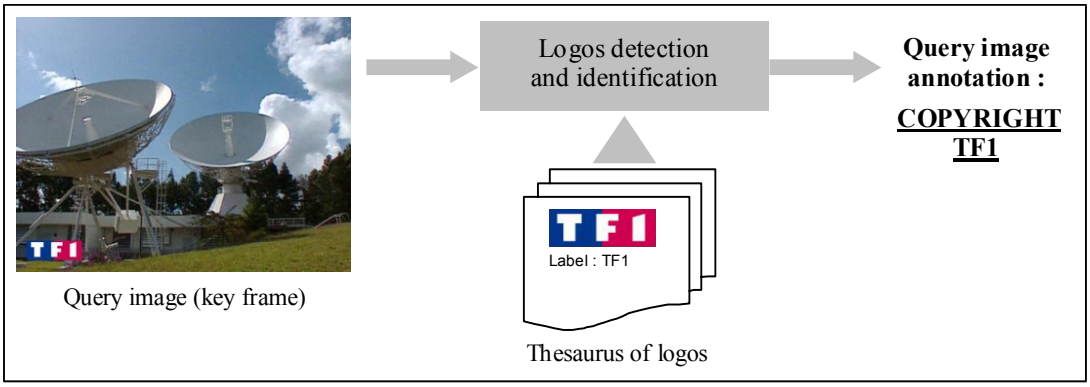


Figure 6: TV logos detection, identification and image annotation principle, with the example of the French TV channel TF1.

More precisely, the approach we present can be decomposed into two phases:

- **Indexation of the thesaurus by visual content (off-line phase).** Given a set of logos that have been annotated by textual information, this phase consists in indexing *the visual content* of each referenced logo. The content-based approach of logo description we propose is presented in section 3.1. It is based on the extraction and the characterization of points of interest. We present these two stages in sections 3.1.1 and 3.1.2.
- **The logos detection, identification and query image annotation (on-line phase).** It consists first in indexing the visual content of the query image by using the same description as the logos one, and second in comparing this description with the ones of the indexed thesaurus. This part can be decomposed into 4 steps that are detailed in section 3.2.

In section 3.3, we illustrate our approach with some representative examples of automatic image annotation from images and logos supplied by a French TV channel. Finally, we discuss the results obtained and list the limits and perspectives that we envisage in section 3.4.

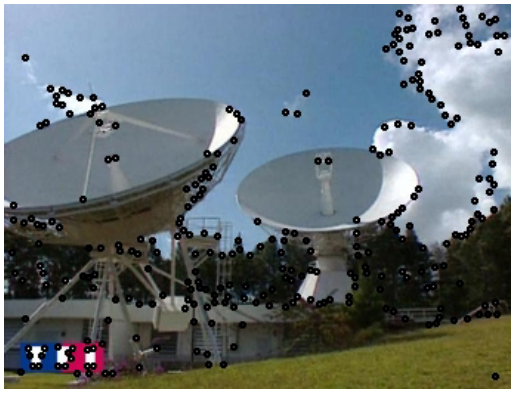
3.1. Logo and image description by visual content

In this paper, we consider that a logo object has the following characteristics: it is a plane object superimposed on the image and occupying a small surface. It can undergo 2D translations, small changes of scale and changes of illumination, from the logos of the thesaurus. In the applications encountered, logos can be superimposed everywhere in the image and multiple logos per image are possible. Due to the multiplicity of logos in the thesaurus, it is not possible to exploit a model of logo, as it is done with faces approaches for example. These constraints impose a *generic* and *local* description of the image. According to these considerations, the approach we choose to implement consists in describing logos and query images by a set of *points of interest*.

In the two next sections, we present the approaches we propose to automatically extract points of interest from color images and characterize them according to the constraints of the application aimed.

3.1.1. Points of interest extraction

When applied to image retrieval, image matching based on points of interest needs points with excellent *repeatability*, i.e. points that can be automatically extracted from images with the same accuracy and under various conditions like viewpoint or illumination changes. Many point extractors exist for gray value images, for example [1] [7] and only one for color images [8]. It has been shown in [9] that it is the color operator which fits better for the required repeatability. This is the one we use to extract points in the whole image during the image characterization step. Examples of interest points extracted on a query image and on the corresponding logo of the thesaurus are presented in Figure 7.



Query image

Corresponding logo
in the thesaurus

Figure 7: Harris Color Points extraction. The points of interest extracted are superimposed in black on the images.

3.1.2. Photometric characterization of the interest points

In a second step, it is necessary to describe the points extracted in a feature space which is function of the photometric information around the point. Many approaches exist for points of interest characterization, see for example [10] [11] [12] [13] [14]. They deal with different image transformation, like 2D translation and rotation, changes of illumination and changes of viewpoint. According to our definition of logo, we have to consider quantities that are invariant to 2D translations, changes of illumination and small changes of scale. Let us consider the *local jet* of the signal at point *(x,y)* computed until order *n*:

$$J_n(x,y,\sigma) = \left\{ I_{i_1 \ldots i_k}(x,y,\sigma) / k = 0,\ldots,n \right\}$$

where $I_{i_1 \ldots i_k}(x,y,\sigma)$ represents the k^{th} derivative of the image channel *I* relative to the $i_1,\ldots,i_k$ variables (*x* and *y*) and σ the size of the Gaussian smoothing applied during the derivatives computation.

This local characterization is computed for each RGB channel of the image. For instance, it gives a set of 18 invariants until order 2. These quantities are invariant to 2D translation. They can easily be made robust to affine illumination changes, by considering either ratios of derivatives or image pre-processing like local image normalization [15]. As it stands, this characterization is only robust to very small changes of scale. We distinguish two ways to make it robust to the small scale changes considered in our application. A first theoretical solution consists in considering scale-space implementations of the local jet, see [7] for example. It makes the characterization invariant to high ranges of scale. In the case of our application where the scale changes are small, we can consider a more simple solution that consists in keeping the original local jet but in adapting the combined similarity measure to deal with these particular *constraint* degrees of freedom. This point is presented in the next paragraph, where we go back over the problem of scale changes.

The similarity measure combined with this characterization is the Mahalanobis distance δ^2, because it takes the different magnitudes and possible correlations of the components into account and includes a model of noise. The covariance matrix associated must be precisely estimated by considering training sequences of representative points from representative images. Such an estimation allows to integrate in the model of noise some interesting aspects:

- Of course, the noise related to the image acquisition (sensors and sampling errors), to numerical errors during the computation of derivatives, to points of interest delocalization, etc;
- Computing the covariance matrix of the features on points sequences differing from constraint scale changes (smaller than 20% for example) allows to fit the characterization to our particular application more precisely than scale-space approaches that do not consider constraint scale changes;

- Computing the covariance matrix on points sequences after normalization to illumination changes (features or image normalization) allows to take the approximations and numerical errors inherent in the normalization into account.

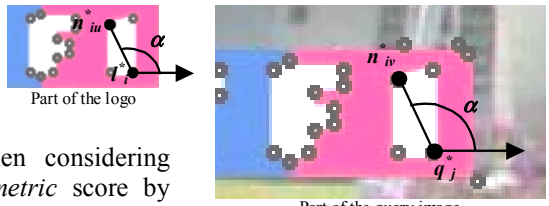
The logo description we propose is based on the local jet and on the Mahalanobis distance characterized by a covariance matrix finely estimated for the particular constraints of our application. The feature space implied will be noted (V, δ^2) in the paper. Its size depends on the order considered for the invariants computation. It represents the space of content-based characterization of the logos and of the query images.

3.2. Logos detection, identification and image annotation

Given a query image that may contain logos and that has been described by using the point characterization presented above, the on-line phase of detection, identification and annotation can be decomposed into 4 steps we describe below:

Step #1: Nearest-neighbor search. This process consists in searching in (V, δ^2) the nearest neighbors of all the points of interest related to the query image. Each point $q_j \in (V, \delta^2)$ of the query image is compared with each point $l_i \in (V, \delta^2)$ of the logos thesaurus. The matches $\{(l_i,q_j)\}$ associated to a distance higher than a given threshold (χ^2 like) are eliminated; the others represent a set of *potential* matches, according to our photometric description. Note that it would be hazardous to keep only the best neighbor of each point of the query images, since the photometric characterization is not exempt from noise and ambiguities.

Step #2: Refining the points matching by adding semi-local constraints of neighborhood. Let us consider a potential match $(l_i,q_j) \in (V, \delta^2)^2$ obtained at the previous step and its respective Cartesian coordinates (l^*_i,q^*_j) in the image plane (the asterisk symbol will denote Cartesian coordinates in the rest of the paper). A basic constraint on neighborhood consists in reinforcing this couple if in the neighborhood of l^*_i there are enough points matched with neighbors of q^*_j. In particular, this constraint allows to eliminate isolated points in the image. Some heuristics based on geometrical constraints can be added to this one. It depends on the geometrical transformations existing between the primitives to match. When considering translation and scale changes, it is possible to compute a *geometric* score by exploiting the property of angles conservation between matches and their neighbors, as illustrated in the figure opposite. For each potential match, a *global* score is computed by combining the photometric score with the geometric one. See [10] for further details on the computation of this score. Finally, we eliminate the matches involving an element which is too involved in another better match (cross matching step). At best one match per query point is kept; they are called *key* matches. See Figure 8 for an example of key matches obtained on the TF1 logo.



Step #3: Estimation of the best transformations parameters. At the end of the previous step, only a few logos of the thesaurus involve key matches; they represent the *candidate* logos for the query image. For each candidate logo, the matches involved $\{(l^*_i,q^*_j)\}$ are used to estimate the geometric transformation T_{LQ}, combining translation plus scale and existing between him and the potential logo of the image query. This estimation requires at least 2 matches to estimate the 2 translation parameters along x and y axis and the scale parameter. In practice, a least square estimation is done on the whole set of matches (or on a subset of best matches).

Step #4: Identification and annotation. At this step, the T_{LQ} transformation estimated for each candidate logo is used to compute the *exact* location $T_{LQ}(l^*_i)$ in the query image of all the points l^*_i characterizing the logo. It allows to build a more precise set of matches than the initial one based on *extracted* points and especially to match possible logo points that were not matched yet, as illustrated on the example of Figure 8. This process supposes that the T_{LQ} have been estimated precisely. Let notice that it is not possible to compute the exact locations directly on all the pixels of the logo image, since some pixels may not belong to the logo (concave form, etc).

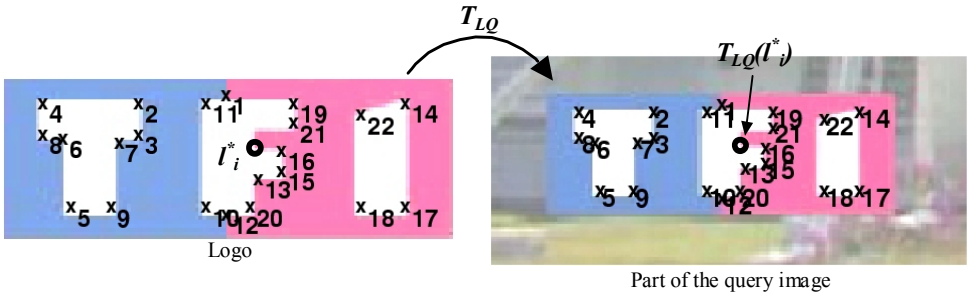


Figure 8: T_{LQ} estimation from key matches (22 couples represented with numbered crosses) and computation of the exact corresponding point location $T_{LQ}(l^*_i)$ of a logo point l^*_i not matched.

A *vote* $V(l_i)$ is then attributed to each point of the candidate logo, that is function of the distance $\delta^2(l_i, T_{LQ}(l_i))$ computed in (V, δ^2):

$$V(l_i) = \begin{cases} 1 & \text{if } \delta(l_i, T_{LQ}(l_i)) < \varepsilon \\ 0 & \text{if } \delta(l_i, T_{LQ}(l_i)) \geq \varepsilon \end{cases}$$

where ε represents a given threshold (χ^2 like). Then we compute a global vote for each candidate logo, by considering the average of all the votes $V(l_i)$ of the points l_i characterizing the logo. This vote represents the proportion of points associated with the logo found in the query image in a similar configuration, thus it indicates its presence or its absence in the image. Finally, the annotation of the query image consists in providing for the best candidate logos (associated to the best votes) its label and if necessary the vote associated, which symbolises a confidence degree concerning the decision.

3.3. Examples of image annotation

We consider a thesaurus of about forty logos of TV channels and several images extracted from television video news. All images (logos and query images) were supplied by the French TV channel TF1. Figure 9 shows some of the most representative logos listed in the thesaurus, annotated with textual labels.

Figure 9: Samples of annotated logos in the thesaurus.

According to this thesaurus, we present in Figure 10 four results of automatic image annotation. For each query image, we give the textual annotation(s) proposed by our approach, with a percentage which represents the vote associated to each candidate logo. Only the candidate logos related to votes higher than 60% are kept. This score traduces that more than 60% of the points of interest characterizing the logo have been matched correctly in the query image. The images by T_{LQ} of the kept logos borders in the query image are superimposed in white on the image to give an idea about the precision of the T_{LQ} estimation.

a) TF1 (99%)

b) From upper-left, clockwise: CNN (85%), CNN (62%), CNN (76%), CNN (92%)

c) No logo found

d) No logo found

Figure 10: Four examples of automatic image annotation. The textual annotations proposed are presented under the images, with the degree of confidence in the decision. The white rectangles drawn represent the images by T_{LQ} of the borders of the corresponding logos in the thesaurus.

The query image a) illustrates the detection and identification of the TF1 logo. The corresponding logo in the thesaurus roughly has the same size. It explains the very good result of the recognition (in terms of confidence rate and of estimation of the T_{LQ} transformation). The query b) shows logo identification in presence of scale changes. Here, to do the test, we had to synthetically superimpose the logos at different scales on the query image. The logo of the thesaurus that has been recognized roughly has the same size as the one on the bottom-left corner of the image, associated with a vote of 92%. We see clearly that the identification inversely decreases with the change of scale. Nevertheless, all the instances of this logo have been correctly identified. The query c) shows a more complex image without logo. All the logo votes computed for this image are smaller than 10%. The query d) presents a small logo on the bottom-left, with transparency. This logo is present in the thesaurus (see samples of Figure 9) but is not recognized, because it has got a small vote (31%) but above all because it is not sufficiently distinctive from the other votes.

3.4. Discussion and conclusions

The method we have presented makes possible to detect the presence of logos in the image, while identifying them, by returning - if they exist - the textual labels(s) related to the logo(s) of the thesaurus closest to part(s) of the image in term of visual similarity. The application considered here is the automatic authentication of image sources and rights for the audio-visual domain. The scenarios presented show that our approach is useful to contribute to automatic characterization of image rights, by proposing labels(s) associated with logo(s) from a thesaurus. In section 3.3, we have shown the performances and the limits of our approach. This particular application implies specific photometric and geometric transformations between models and query images. Here, we considered 2D translations, small changes of scale and changes of illumination. If we want to deal with higher changes of scale, a simple solution consists in considering logo models at different representative scales in the thesaurus. The example d) of Figure 10 shows the limits of our point characterization approach with the problem of transparency that sometimes occurs with logos.

Other applications of automatic sub-image identification are possible with our approach. Let us consider for instance the automatic detection and identification of trademarks in video sequences. As illustrated in Figure 11, the Esso logo involved is not in the plane of the camera and probably is not plane. Here, we see that the points of interest approach can be very pertinent, since it allows to find partially occulted logos. Of course, a different photometric and geometric characterization adapted to the transformations involved must be applied. In the example of Figure 11, we used a point description that is locally robust to changes of viewpoint. See [10] for more details on this version.

Figure 11 : Automatic detection and identification of the Esso trademark in key frames of a video (1 image per second). The images involving the searched logo are presented by decreasing order of similarity (from top-left to bottom-right).

CONCLUSION

In this paper, we have presented a hybrid thesaurus (textual and visual) approach for semantic object recognition and identification for video news archives. Both faces and TV channels logos were concerned for this automatic annotation. Image captions or audio transcriptions are not always available to extract the requested semantic information such as persons or TV channel names. Completely image-analysis-based methods are presented for face detection, close-up estimation, person (VIP) identification, logo detection and recognition. This thesaurus approach is very suited to provide archivists online central and updated references for most frequently encountered semantic video objects.

REFERENCES

1. R. Brunelli and T. Poggio. *Face recognition: Features versus templates*. In Pattern Analysis and Machine Intelligence. vol (15)10, 1042-1052, 1993.
2. Wiskott, L. and Fellous, J.M. and Kr, ger, N. and Malsburg, C. *Face Recognition by Elastic Bunch Graph Matching*. In International Conference on Computer Analysis of Images and Patterns, 1997.
3. S. Gilles, *Robust Description and Matching of Images*. PhD thesis, Oxford University. Oxford University, 1998.
4. R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.
5. Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade *Name-It: Naming and Detecting Faces in News Videos*, IEEE MultiMedia, Vol. 6, No. 1, January-March, pp. 22-35, 1999
6. C. Harris and M. Stephens, *A combined corner and edge detector*, In Alvey Vision Conference, pp. 147-151, 1988.
7. K. Mikolajczyk and C. Schmid, *Indexing based on scale invariant interest points*, In ICCV, pp. 525-531, 2001.
8. P. Montesinos, V. Gouet and R. Deriche, *Differential Invariants for Color Images*, in IAPR ICPR, Australia, 1998.
9. V. Gouet and P. Montesinos, R. Deriche, D. Pele, *Evaluation de detecteurs de points díinteret pour la couleur*, RFIA 2000.
10. V. Gouet and N. Boujemaa, *Object-based queries using color points of interest*, IEEE Workshop CBAIVL/CVPR Hawaii, 2001.
11. V. Gouet and N. Boujemaa, *On the robustness of color points of interest for image retrieval*, IEEE-ICIP, Rochester, 2002.
12. A. Baumberg, *Reliable feature matching across widely separated views*, In CVPR, PP. 774-781, 2000.
13. F. Schaffalitzky and A. Zisserman, *Multi-view matching for unordered image sets*, in ECCV, pp. 414-431, 2002.
14. C. Schmid and R. Mohr, *Local gray value invariants for image retrieval*, PAMI, 19(5):530-534, 1997.
15. V. Gouet and P. Montesinos, *Normalisation des images en couleur face aux changements díillumination*, Angers, RFIA 2002.