GENERIC OBJECT TRACKING FOR FAST VIDEO ANNOTATION

Rémi Trichet and Bernard Mérialdo
Institute Eurecom, BP 193, 06904 Sophia-Antipolis, France

Keywords: Object tracking, video annotation, keypoints, interactive television, color adaptation.

Abstract: This article describes a method for fast video annotation using an object tracking technique. This work is part of the development of a system for interactive television, where video objects have to be identified in the video program. This environment puts specific requirements on the object tracking technique. We propose to use a generic technique based on keypoints. We describe three contributions in order to best satisfy those requirements: a model for a broader temporal use of the keypoints, an ambient color adaptation pre-treatment enhancing the keypoint detector performance, and a motion based bounding box repositioning algorithm. Finally, we present experimental results to validate those contributions.

1 INTRODUCTION

Object tracking has been the subject of recent improvements (Isard, 2001) (Comaniciu, 2002) (Hu, 2004) (Techmer, 2001) (Gyaourova, 2003) (Pupilli, 2005), so that these systems are becoming more and more reliable. In consequence, their application for commercial products, such as video surveillance, is becoming more and more common. But those trackers are often developed for specific applications, leading them to solely work in a very constrained environment. For instance, surveillance systems (Comaniciu, 2002) (Techmer, 2001) and traffic monitoring (Gyaourova, 2003) make the assumption of a pre-defined pattern of object and a still camera. Body trackers (Isard, 2001) (Karaulova, 2000) use specialized model like articulated stick figure to represent human beings. Moreover, in most of the human trackers (especially in the surveillance domain) only indoors places are considered. In sport applications (Jaffré, 2003), the tracked object is always small and fast with strong color contrast.

Our work takes place in the context of the development of an interactive television system, which aims to realize direct interactivity with moving objects on hand-held receivers. In this system, the video producer will annotate the video program by defining video objects in the video sequence, and attaching to them some additional content (for example, text, images, videos, web reference, etc…). On the receiver side, the user watching the video program will be able to select the active video objects and immediately access the corresponding additional content. This environment induces several specific constraints:

- For this scheme to be practically viable, the extra production cost for manual annotation should be minimized, so manual annotation should be as fast as possible. The idea is that the producer will identify a video object on the first frame of a shot, and a tracking system will follow it throughout the shot. Since the producer has to check the validity of the tracking, the tracking system should be as fast as possible (and if possible faster than real time).
- This system could be used for any type of video program, so the tracking system should not be constraint on a specific video genre, or specific characteristics of the video object. It should rely on generic techniques, or be able to adapt to the video program being annotated.
- Both for annotation and display to the user, the video objects should be identified by bounding boxes. This limits the refinement of the description of the video objects, and imposes the correct placement of the bounding box as a performance criterion for the tracking system.

Our approach is based on the identification of keypoints in the video scene. Keypoints offers many advantages. These points, first developed for robotic

Trichet R. and Mérialdo B. (2007).
GENERIC OBJECT TRACKING FOR FAST VIDEO ANNOTATION.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 419-426

(Moravec, 1980), are located at key positions (usually corner or extrema of a given function), making them easy to recover. Moreover, they are enriched by local descriptors in order to increase their robustness to usual transformations (scale changes, illumination changes, rotations, affine transformations…). Thus, keypoints are a reliable tool for the problem of generic tracking. Moreover, their computation is independent of the object location, so that most of the computation and the matching process could be done offline, leaving only minimal processing left during the annotation session.

We propose three contributions to satisfy the specific requirements of our video annotation environment: first, a model for a broader temporal use of the keypoints; then, an ambient color adaptation pre-processing allowing the keypoint detector to deal with a larger variety of videos; and finally, a motion based bounding box repositioning algorithm.

The rest of this article is organized as follows. We will describe the structure of our tracking system in section 2. Then we motivate the choice of our keypoint detector in section 3. The color adaptation pre-processing is explained in section 4. Section 5 describes our multiframe model. In section 6, our bounding box repositioning algorithm is discussed and results are presented in section 7. Finally, section 8 will conclude and suggest further enhancements.

2 STRUCTURE OF THE TRACKER

Our tracking system is modelling the object with a set of keypoints. The model is initialized with the keypoints that lie within the bounding box of the object in the first frame. Afterwards, the keypoints extracted on every new frame are matched with the model keypoints using a winner-take-all algorithm. The bounding box is then repositioned according to the motion of the matched points. Finally, the model is updated: the descriptors of the model matched points are updated with their corresponding image's point descriptor, and the new object keypoints are added to the model. This process can be summarized by the following steps:

Initialization:
- keypoint extraction for the first frame (off-line)
- object bounding box drawing

Main loop (for every new frame):
- keypoint extraction (off-line)
- keypoint matching (off-line)
- bounding box repositioning
- object model update

When annotating a recorded video program, it is important to notice that much of the required computation can be performed before the annotation session: keypoints may be computed over the whole image for every frame (we don't know yet what are the objects that will be annotated), and matched between consecutive frames. Those results can be stored, so that, during the annotation session, the only work left is to reposition the bounding box and update the object model for every new frame. This organization of the computation answers to the requirement for a fast tracking system.

3 KEYPOINTS DETECTOR

The invariant feature detectors can be divided in two main categories: keypoint detectors and regions detectors. Although they both rely on the same principles, some characteristics differ. Regions detectors, by extracting larger areas, have a less precise localization. On the other hand, the associated support region used for the descriptor computation will be the extracted region, instead of a pre-defined circular region. This much more accurate region will lead to a more robust object description. In order for the reader to understand our reasoning in the choice of a keypoint rather than a region detector, this section will address feature detectors.

There are two criterions for establishing the efficiency of a feature detector. The first criterion, called repeatability rate, was set up by (Schmid, 1997). It represents the ability for a point to be detected at the same location in various images. In practice a point is rarely detected at the exact same position, but in a small neighbourhood. This implies the notion of precision of the localization, introduced by (Gouet, 2000). In her work, she specifies that this phenomenon is independent from the repeatability rate.

The second criterion is the robustness to standard transformations (rotation, blur, scale changes, affine transformations…). This criterion is related to the repeatability rate in the sense that a transformation unhandled by a detector will lead to a low repeatability rate. Hence, the repeatability rate is usually computed for a given transformation. The evolution of the detectors could be associated to their ability to deal with these transformations. At the infancy of the domain, the algorithms solely rotation were handled. The widely used Harris corner detector (Harris, 1988) was giving the better score. Latter, (Dufournaud, 2000), followed by (Mikolajczyk, 2001) with the Harris-Laplace detector have strengthen the Harris detector under the scale changes. Finally, the principle was recently extended to allow the algorithms to deal with affine transformations. A survey and comparison of the affine invariants algorithms is shown in (Mikolajczyk, 2005-2).

A large variety of descriptors has been developed or adapted to the feature detectors in order to enhance their discriminative power. A quite exhaustive survey and evaluation of these descriptors is proposed in (Mikolajczyk, 2005-1).

In the framework of a video tracker, the repeatability and the precision of the localization of our detector will be our first priority. Moreover, our points will often be located in areas of frequent modifications (a moving object). Thus, descriptors computed on a small size region will be preferred. These two cues have oriented our choice on keypoint rather than region detectors. Furthermore, only slight changes will happen between two frames. So, our system will mainly have to deal with blur, rotation, and some scale changes.

The Harris-Laplace detector (Mikolajczyk, 2001) satisfies all these requirements and was implemented for our experiments described in the others sections.

The associated descriptors are the generalized color moments (Mindru, 2003). They are an adapted version of the grey-values moments to the color channels. A generalized moment of order *p+q* and degree *a+b+c* is defined by:

$$M_{pq}^{abc} = \sum_{V} x^p y^q \left[R(x,y)\right]^a \left[G(x,y)\right]^b \left[B(x,y)\right]^c \qquad (1)$$

They characterized the shape of the distribution on a region in a uniform manner. Moreover the exploitation of color channels allows them to extract more information. So, they can describe more precisely a region without needing the computation of higher degrees moments. This discriminative power is more widely exploited by our pre-treatment mechanism presented in the fifth part. The support region chosen is a circular region with a radius proportional to the detected scale.

4 COLOR ADAPTATION

Our tracking system should handle a diversity of video genres, in particular a diversity of color ambiance. Thus, we propose a color adaptation scheme to weight the discriminative importance of different color channels. For instance, in the figure 1, we have examples of several images where one color is predominant. When the information is concentrated in one or two color channels, assigning to each channel an importance proportional to its discrimination power will extract richer information and will help the detection algorithm to produce better results. In our model, we try to associate to each color channel *c* a weight *P(c)* representing his importance, so that:

$$\sum_{c=1}^{n} P(c) = n \qquad (2)$$

with *n* being equal to the number of channels (3 in our case of RGB color space).

For comparing the relative importance of color channels, we have defined two indicators : size and saliency. The size represents the extent and the intensity of the color channel in the image. The size *S(ci)* of a channel *c* is equal to the sum of the pixel intensities of the image *Im* for this color channel:

$$S(c) = \sum_{p \in \mathrm{Im}} I_c(p) \qquad (3)$$

The saliency models the visual attraction of a particular color channel. It is characterized by a pronounced histogram on one or several values (presence of peaks). We determine it by calculating the Kurtosis for each of the channels. The Kurtosis *K(c)*, or central moment of degree 4 represents the flattering level of a distribution. A Kurtosis lower than zero indicates a flat distribution, whereas a Kurtosis higher than zero characterizes a peaked distribution. More formally:

$$K(c) = \frac{\sum_{x=0}^{n} (x-\mu_c)^4 h(x)}{\left(\sum_{x=0}^{n} (x-\mu_c)^2 h(x)\right)^2} - 3 \qquad (4)$$

with *n* the number of bins (255 in the case of the RGB color space) of the considered histogram, *h(x)* the xth value of the distribution, and μ_c the distribution average. The weights associated to each channel are then obtained by combining the size and the saliency with the following formula:

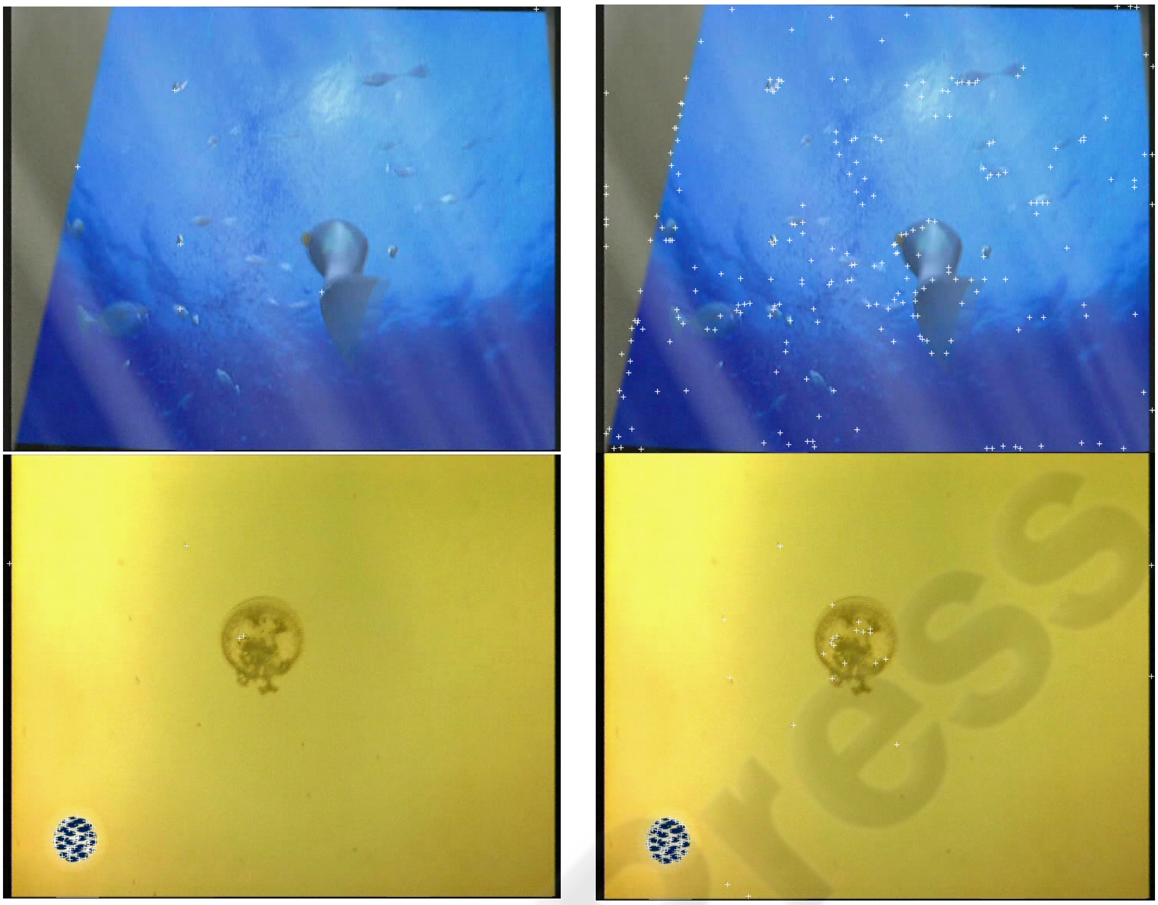

Figure 1: Harris keypoints extracted (represented by a white cross) in the sequences “fish” and “jellyfish” (a) without color adaptation to the importance of the channels (b) with color adaptation.

$$P(c) = S(c) \times (1 + K(c)) \qquad (5)$$

Finally, the weights P(c) are normalized to satisfy the initial constraint (2). Figure 1 provides examples of the results of the Harris detector with and without color adaptation.

5 MULTIFRAME MATCHING

Our matching algorithm associates to each point of the previous frame the most similar point of the current frame in a local neighborhood. The similarity between points is based on the Mahalanobis distance of their descriptors. Each point of the current frame is only associated once with a winner-take-all like algorithm, that is, the best matching is always preferred and no global matching quality is considered.

As explained in the section 3, the efficiency of keypoint detectors could be evaluated with their robustness to usual transformations and their repeatability. However, this last measure has rarely been used in the context of videos, but rather for images. In our experiments, we have observed a temporal instability of the keypoints in videos. Because of some local alterations specifics to videos (blur due to motion, poor quality video, or sensors calibration problems), some points disappear during one or several frames, and then reappear. To overcome this drawback, our matching algorithm is conserving the keypoints during *k* frames, that is, if a keypoint is not matched for *k* images, then it is removed. Our experiments show an increase of the matching rate when the keypoint conservation time increases (see Figure 2). But when a keypoint is conserved for too long, its descriptor is not updated and this increases the risk of false matching and does not lead to a better accuracy in tracking (see figure 3). We have chosen *k=3*, since this offers a good compromise between tracking accuracy and computation time.

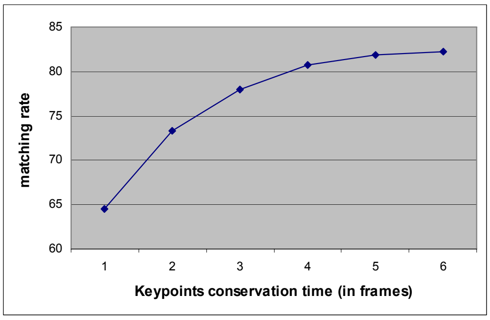

Figure 2: Matching rate function of the keypoint conservation time.

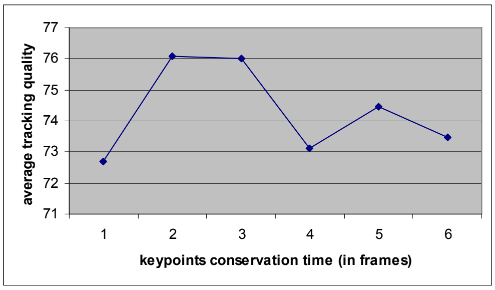

Figure 3: Average tracking quality function of the keypoint conservation time. The tracking quality is computed with the mechanism discussed in section 7.

6 BOUNDING BOX REPOSITIONING

Only few tracking algorithms are using keypoints. Gabriel & al (Gabriel, 2005) work is the most striking one in this particular field of research. This algorithm takes as input the matched keypoints of the observed image. It first calculates a trimmed mean over the motion values, considering the upper and lower bounds as false matching. The gravity centre of these points is then computed for the model and the observed image. In order to eliminate more false matching, the remaining set of points is then passed through another filter. If erasing a point moves significantly the position of the gravity centre, then this point is considered as a false matching and will not be further taken into account. The last step consists in positioning the bounding box on the observed image in order to have the gravity centre at the same position as for the model. This method has two drawbacks. First, scale changes are not handled. And secondly, because of the aforementioned temporal instability of the keypoints (see section 3), points could appear and disappear from one image to another, modifying the position of the gravity centre. This problem, added to the keypoint localization imprecision (mentioned in sectioned 3 too) will reduce the precision of the gravity centre detection. Since the bounding box positioning is done in relation to the previous image, errors will propagate during the sequence.

In order to avoid most of theses disadvantages, we have chosen a motion based positioning algorithm. Knowing the position *(x,y)* of the keypoints for the object model *A* and the observed image *B*, we use the least square method to compute the translation values *a0* and *a1*, the scale values *a2* and *a3*, and the rotation values *a4* and *a5* that best explain the motion between *A* and *B*. More formally, we have:

$$\begin{pmatrix} x_A \\ y_A \end{pmatrix} = \begin{pmatrix} a2 & a4 \\ a5 & a3 \end{pmatrix} \times \begin{pmatrix} x_B \\ y_B \end{pmatrix} + \begin{pmatrix} a0 \\ a1 \end{pmatrix} \tag{6}$$

In order to facilitate the calculation of a local movement, we have fixed *a4=a5=0*, we first identify the translation parameters, then we explain the remaining error with the scale parameters. As for Gabriel's method, we eliminate keypoints which are potential false matching by the condition:

$$(m_x - x) < k\sigma_x \quad et \quad (m_y - y) < k\sigma_y \tag{7}$$

where m_x and m_y are the means, σ_x and σ_y and the standard deviations of the keypoints coordinates on the *X* and *Y* axes respectively, and k the tolerance factor of motion (fixed to *k=2*).

7 EXPERIMENTAL RESULTS

To evaluate the improvement brought by the contributions proposed in our approach, we have set up an evaluation testbed. In order to test the genericity of our algorithm and to compare the algorithms in different situations, videos sequences with various objects and difficulties have been chosen. These shots are described in table 1. A hand labelled ground truth bounding box was set up for each frame and accuracy of the tracking algorithms hypothesis is measured by the following classical formula:

$$\mathrm{d(A,B)} = \frac{A \cap B}{A \cup B} \tag{8}$$

Table 1: Shots description.

Video Name	Object Size	Difficulties	Description	Frames
Fashion	Big	None	woman turning back	120
Soccer	small	None	Football player tracking	70
Cooking	medium	scale change, cluttered background	Marmite tracking with camera movement	60
Cognac	small	Occlusions, fast & irregular motion, cluttered background	Cognac bottle tracking	30
Jellyfish	medium	low contrast, fast object change	jellyfish swimming	30
Frying pan	medium	cluttered background	Cook showing a frying pan	75
Bottle	small	Occlusions, cluttered background	bottle passing from hand to hand	60

Table 2: Comparison of our algorithm based on a motion model, Gabriel's algorithm based on the centre of gravity position, and the basic Meanshift. For a given frame, the number displayed is the average performance over all the previous frames. Best results are highlighted in yellow.

Video Name	Frame	Our algorithm	Gabriel's algorithm	Basic MeanShift	Our algorithm without color adaptation
Fashion	30	89,5742%	89,4678%	75,30115%	89,2855%
	60	79,1712%	83,1872%	66,92%	79,0293%
	90	77,5669%	79,5691%	62,8489%	76,573%
	120	78,3963%	78,9066%	62,4674%	78,6172%
Soccer	30	70,6246%	68,7984%	84,9789%	73,7586%
	70	76,5633%	59,1253%	81,8645%	72,4328%
Cooking	30	90,4768%	85,1282%	72,9395%	92,6239%
	60	75,7842%	64,6829%	68,7116%	80,6439%
Cognac	15	72,4329%	71,8726%	60,8272%	78,4865%
	30	53,2051%	47,5774%	40,5432%	45,1862%
Jellyfish	15	65,0661%	85,6222%	86,4634%	58,123%
	30	52,7346%	70,9549%	84,5328%	42,458%
Frying pan	25	92,1674%	72,9197%	78,7184%	83,6158%
	50	81,8894%	52,484%	69,4585%	76,4032%
	75	79,8878%	46,0843%	63,9549%	80,5012%
Bottle	20	96,7539%	96,3405%	54,17%	96,7539%
	40	73,6986%	67,7005%	47,2221%	70,6699%
	60	62,3459%	55,0777%	37,2699%	59,5623%
Average quality		76,02%	70,86%	66,62%	74,23%

where *A* and *B* are respectively the ground truth and the method bounding boxes. The results in table 2 show the average performance of the tracking for different video sequences, and various intermediate frames. A steep decrease in the performance is generally the sign of a temporary loss of the object.

We have compared the performance of our algorithm with Gabriel's (Gabriel, 2005) and the basic Meanshift (Comaniciu, 2002). Gabriel's method was implemented with the same keypoints as ours, and using our color adaptation mechanism, so that only the bounding box repositioning algorithms are compared. The basic Meanshift was used without the color adaptation algorithm.

Results are presented in table 2 and some examples are shown in figure 4. They show a better behaviour than Gabriel's algorithm in most of the cases except when the whole object has local motion variations, as the "jellyfish" sequence shows. However our algorithm is still sensitive to error propagation during the shot. In consequence, in the case of a temporally loss, our algorithm will have problems to focus on the object again. This problem could dramatically decrease the performance of the tracker in the case of a long sequence. Our algorithm usually outperforms the basic Meanshift except for small or medium objects in an uncluttered environment (see the "jellyfish" and the "soccer" sequences scores), cases in which this method is known to behave well. The Meanshift quite low results could be explained by a lack of precision. Actually, we observed that, except for the "bottle" and the "cognac" shots where the object is lost, the algorithm manages to focus on the object, but oscillates around it.

Table 2 also shows the results for our algorithm without the color adaptation mechanism. The slightly better scores could not be considered as significant because of the subjectivity of the ground truth. Nevertheless, in the case of a pronounced color, the “jellyfish” video for instance, a real improvement is stated.

8 CONCLUSION

We have presented our work in developing a generic object tracker for fast video annotation based on keypoint detection. The video annotation environment imposes specific constraints on the characteristics of the object tracking, and this lead us to propose three contributions the tracking: an ambient color adaptation mechanism, a matching algorithm with a temporal use of the keypoints, and a bounding box repositioning algorithm based on a motion model. All these enhancements were validated through an evaluation testbed composed with video sequences including various difficulties.

But some flaws still remain, notably the fact that errors propagate through the sequence. To overcome this problem, we would like to label each points “object” or “background”. These labels will further be used, to enhance the bounding box repositioning algorithm by maximising the number of “object” points inside the bounding box and minimizing the “background” ones. A probabilistic matching algorithm using the point’s neighbourhood relations is also being studied.

REFERENCES

Comaniciu D., Meer P., 2002, Mean Shift: A Robust Approach Toward Feature Space Analysis, *IEEE Trans. Pattern Anal. Mach. Intell. 24(5): 603-619.*

Dufournaud Y., Schmid C., Horaud R., June 2000, Matching Images with Different Resolutions, *International Conference on Computer Vision & Pattern Recognition.*

Gabriel P., Hayet J.-B., Piater J., Verly J., 2005, Object Tracking Using Color Interest Points, *in Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS'05).*

Gouet V., Oct 2000, Mise en correspondance d'images en couleur - Application à la synthèse de vues intermédiaires, *Thèse de doctorat, Université de Montpellier II.*

Gyaourova A., Kamath C., and Cheung S.-C., October 2003, Block matching for object tracking, *LLNL Technical report,. UCRL-TR-200271.*

Harris C., Stephens M.J., 1988, A combined corner and edge detector, *In Alvey vision conference, pp147-152.*

Hu W., Tan T., Wang L., Aug 2004, M. S, A survey on visual surveillance of object motion and behaviors, *IEEE Transactions on Systems, Man and Cybernetics, Part C, Vol. 34, No 3, pp. 334- 352.*

Isard M. and MacCormick J., 2001, BraMBLe: A Bayesian Multiple-Blob Tracker *Proc Int. Conf. Computer Vision, vol. 2, 34-41.*

Jaffré G., Crouzil A, 2003, Non-rigid object localization from color model using mean shift, *ICIP (3), 317-320.*

Karaulova IA, Hall P, Marshall A., 2000, A hierarchical model of dynamics for tracking people with a single video camera*. In: Mirmehdi M, Thomas B, editors. Proceedings of the Eleventh British Machine Vision Conference (BMVC2000), p. 352--61.* Bristol: ILES Press.

Moravec, H.P, 1980, Obstacle avoidance and navigation in the real world by a seeing robot rover, *Tech. Rept, CMU-RI-TR-3, The Robotic Institute, Carnegie-Mellon University, Pittsburgh, PA.*

Mikolajczyk K., Schmid C., May 2001, Indexation à l'aide de points d'intérêt invariants à l'échelle *Journées ORASIS GDR-PRC Communication Homme-Machine.*

Mikolajczyk K., Schmid C., 2005-1, A performance evaluation of local descriptors, *IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 27, Number 10.*

Mikolajczyk K., Tuytelaars T., Schmid C., Zisserman A., Matas J., Schaffalitzky F., Kadir F., Van Gool L., 2005-2, A comparison of affine region detectors, *International Journal of Computer Vision, Volume 65, Number ½.*

Mindru F., Tuytelaars T., Van Gool L., Jul.2003, Moment Invariants for Recognition under Changing Viewpoint and Illumination*, Theo Moons,ACM.*

Montesinos P., Gouet V., Deriche R., 1998, Differential invariants for color images*, International conference on pattern recognition.*

Pupilli, M., and Calway, A., 2005, Real-Time Camera Tracking Using a Particle Filter, *In Proceedings of the British Machine Vision Conference, BMVA Press.*

Schmid C. and Mohr R., 1997, Local Greyvalue Invariants for Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Techmer A., 2001, Contour-based motion estimation and object tracking for real-time applications. *In International Conference on Image Processing, volume 3, pages 648--651, Thessaloniki, Greece, 87.*

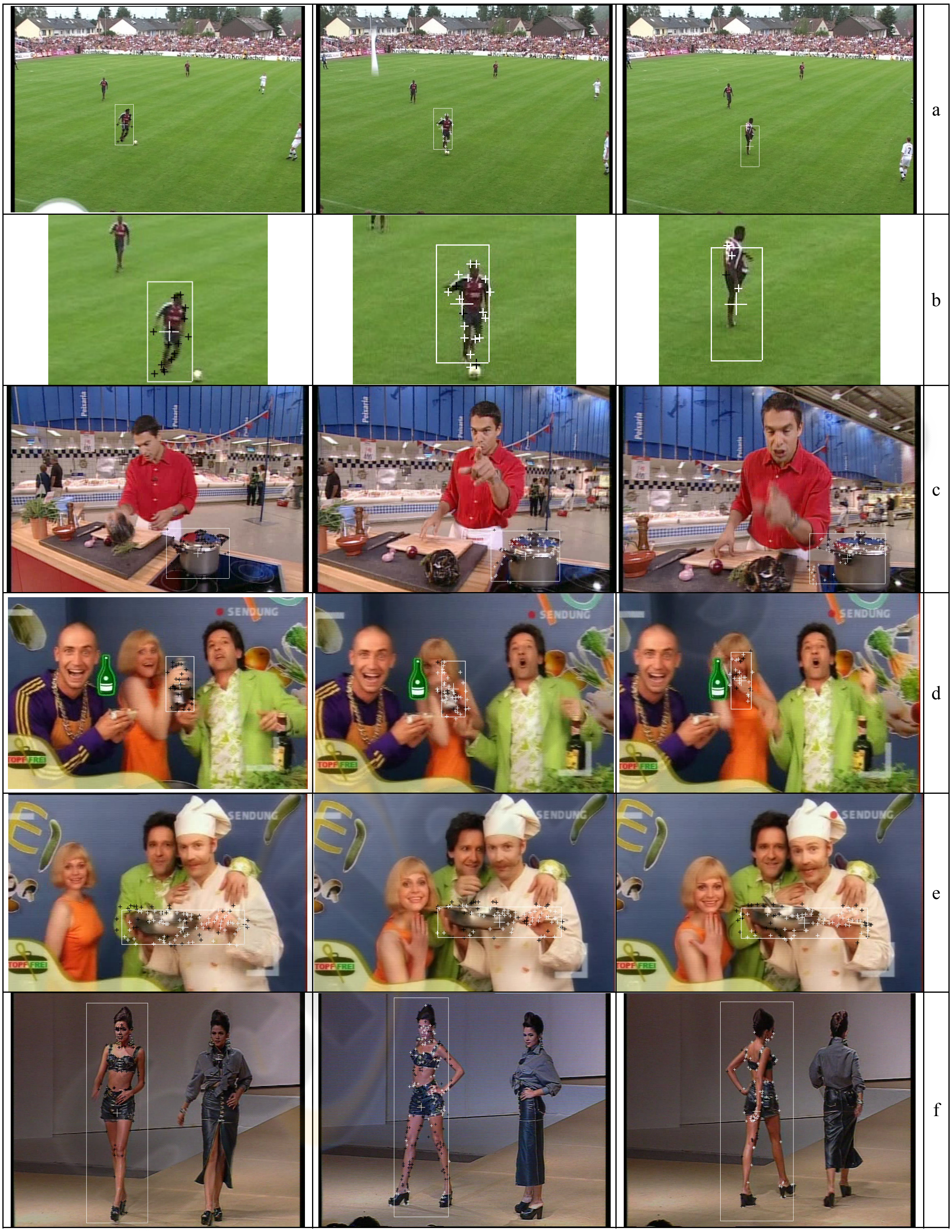

Figure 4: Tracking examples of our algorithm on several videos. Matched keypoints are in white, unmatched ones in black (a-b) “soccer” sequence and the corresponding zoom for the frames 0, 30, 70 (c) “cooking” sequence for the frames 0, 30,60 (d) “cognac sequence for the frames 0, 10, 15 (e) “frying pan” sequence for the frames 25, 50, 75 (f) frames 0,60,120 for the “fashion” sequence.