KEYPOINTS LABELING FOR BACKGROUNG SUBSTRACTION IN TRACKING APPLICATIONS

Rémi Trichet, Bernard Mérialdo
Institut Eurecom, BP 193, 06904 Sophia Antipolis, France
{remi.trichet, Bernard.Merialdo}@eurecom.fr

ABSTRACT

This paper studies the problem of background / object differentiation in a keypoint-based tracking application where the object is delimited with a bounding box. We present a keypoint labeling algorithm based on four features: the label of the matched keypoint, color, motion, and position. We discuss methods to best exploit these features, then we detail our labeling algorithm and validate it with some experiments on several tracking video sequences.

Index Terms— object tracking, keypoint labeling, background subtraction, data association.

1. INTRODUCTION

In order to properly localize objects from one image to the next, tracking algorithms have to define, from a reference frame *t*, the representative object area where the necessary information for the object identification at frame *t+1* will be extracted from. If this area is a precise mask, perfectly delimitating the object, the whole information obtained will be reliable. However, the methods providing such a frontier, like segmentation [1], "snakes" [2], or mesh-based techniques [3] are not reliable or solely work in constrained environment (highly distinguishable contour, strong contrast between object and background…). That is the reason why most of the existing methods rely on a coarse localization of the object, usually with a bounding box. The bounding box will contain both object and background pixels so that features computed from the bounding box will contain a mixture of information. The background pixels are called *outliers*. A large amount of outliers is likely to introduce a big quantity of false information and degrade the performance of the tracking algorithm.

Parameter estimation algorithms [4,5] are usually able to deal with a certain quantity of outliers. Motion can sometimes be used to distinguish the object from the background. However, in the case of object tracking, especially for deformable objects, motion can be subject to important internal variations, and it is then difficult to differentiate, the object from the background motion models.

To summarize, outliers remain an open problem in the object tracking domain. They are often treated as noise by the algorithms, and little work has been done in order to limit their influence. A common trick consist in using an elliptical bounding box rather than a rectangular one, closer to most of the objects form. But this remains often insufficient.

The present work is undertaken in the framework of the Portivity project. This project aims at developing an interactive television system, which can realize direct interactivity with moving objects on hand-held receivers. Fast annotation of moving objects in videos is an essential component of this system, thereby motivating the need of an efficient generic tracking system, able to deal with various genre of videos and objects. Moreover, to facilitate manual annotation, the object is simply defined with a bounding box. In our previous work [6,7], a keypoint based tracking system was built up in order to fulfill the specific requirements of this application. In this particular framework, the keypoints and their corresponding descriptors are extracted in a preprocessing step.

In the specific context of a keypoint-based tracker, tackling the background subtraction problem consists in identifying object keypoints from background keypoints. The idea is similar to data association [8]. We label each keypoint as "object" or "background ". Only the "object" keypoints will latter be used to estimate the object motion. A basic labeling algorithm will consider that keypoints inside of the bounding box are "object" and those outside are "background". To improve the tracking performance, we refine this process by computing for each keypoint the likelihood to belong to the object or the scenery. This likelihood is based on a combination of several features: the label of the matched keypoint, color, motion, and position.

The rest of this article is organized in four sections. First, we detail principles to exploit each of the four features. Then we present our labeling algorithm, followed by experimental results. Finally, we summarize our findings in the conclusion and suggest future improvements.

2. FEATURE ANALYSIS

The keypoint-based tracking is performed in three steps. First, all the keypoints between successive frames *t* and *t+1* are matched according to their corresponding descriptors

[8]. Then, the global object motion is computed and the object is repositioned [7]. Finally, the object model is updated by adding the new keypoints and deleting the obsolete ones [7]. The labeling takes place after the matching step.

The labeling problem can be stated as follows: given two sets of keypoints A and B, respectively coming from successive frames t and $t+1$, a matching between the elements of A and B, and the labels of keypoints in A, what are the labels of the keypoints in B? This means identifying the object points B_O and the scenery points B_B. We introduce the likelihood L_p of a point p as its probability ($0 \leq L_p \leq 1$) to belong to the object. Each point with a likelihood greater to 0.5 will be considered as part of the object, others will be considered as part of the background.

It is important to notice that the bounding box motion will solely be assessed with the matched keypoints. The label allocation of the other points could hence be postponed if we don't have enough information (without matching, the motion information is not available) to reach a reliable decision. In consequence, all the A labels won't be necessarily available, and all the B keypoints won't be inevitably labeled.

These two sets B_O and B_B will be determined in relation with their homogeneity of certain features. The choice and the combination of these features, as well as the way to evaluate their homogeneity (local or global) are crucial on the classification saliency. We have investigated some cues likely to serve this goal. We present in this section their interest as well as the tested techniques.

Matched point label: If a point is matched and his correspondence is already labeled, the likelihood that the label will be the same is very high. Two associated points will only have a different label in the case of a false matching. In this case, assigning the label with certainty would propagate the error. This raises the question of the partial or total influence of the matching on the labeling decision.

Color: The color information is easy to exploit and often discriminatory. It is thus a privileged choice. Moreover, in a keypoint based tracking framework, a descriptor is already associated to each point. Two variants are then possible: The global clustering of the keypoints in two sets according to their color descriptors or the independent classification of each point by comparison with the k nearest "object" label neighbors and the k nearest "background" label neighbors. Our experiments have rapidly shown that the first possibility has to be banished. Indeed, it is based on the assumption of an object and background color homogeneity. But they could be constituted of several feature colors. Moreover, one of the object colors could be more similar to the background than to the rest of the object, leading to misclassification. We have, hence chosen a local color based evaluation of the similarity. After the descriptor comparisons of the k nearest "object" and "background" keypoints, two similarity values S_O and S_B are obtained. The estimated probability P_C of the keypoint to be part of the object according to the color is then given by the formula:

$$P_C = S_O/(S_O+S_B)$$

The k parameterization is a compromise between the amount of gathered information and its reliability. Indeed, a high value of k means a high number of keypoints used for the measure. However, the object could be constituted of many colors and to be solely locally homogenous. So, if k is too large, some of the used keypoints will be too far away from the assessed keypoint and their information may not be discriminative.
We have fixed k to 3 in order to balance these two influences.

Motion: By definition, the object motion differs from the background one. For this feature, we have also the choice between local or global consistency. The principle of a local consistency is, as for color, based on the comparison with the k nearest neighbors for each label. But the keypoint localization is generally not accurate enough to provide a reliable estimation when only a small number of points are used. As a consequence, we prefer a global motion assessment: we find two motion vectors, one for the object and one for the background. The probability for a keypoint to belong to the object or the scenery, regarding its displacement, will be a function of its similarity with these two vectors. The object motion is already computed for each frame in the tracking algorithm. To evaluate the background motion, we note that there is no need for a high accuracy, since the goal is only to distinguish object from background. So, we consider the background motion as a translation, that is calculated as a mean value from the labeled keypoints, after elimination of noise. If m_O, m_B, and m_p are the respective object, background and point motions, the likelihood P_M for the keypoint to be part of the object is given by:

$$P_M = \left(P_M^X + P_M^Y\right)/2$$

$$P_M^a = \mathrm{abs}(m_O - m_p)/(\mathrm{abs}(m_O - m_p) + \mathrm{abs}(m_B - m_p))$$

With a the considered axis and *abs()* the absolute value function.

Position: The keypoint position has revealed being a determinative feature for the label evaluation. Our system is grounded on the hypothesis that the keypoint label is only uncertain for a small portion of the keypoints. The points widely in the interior of the bounding box are considered as object keypoints in proportion with their proximity of the center. On the same way, the more keypoints are distant from the bounding box border, the more they are deemed

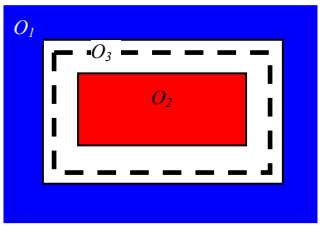

Figure 1: keypoint labeling in relation to their position. The dashed line is the bounding box border. In blue and red, the areas O_1 and O_2 where the keypoints are respectively considered as background and object. Only the white area O_3 keypoints remain uncertain.

to be part of the background. Then, only a small belt in the vicinity of the boundary will be considered (see Figure 1). The point position being not discriminative in this area, the label evaluation is let to some other criteria. The definition of these three areas is crucial for the good behavior of the algorithm because it enables to induce two initial sets of "object" and "background" keypoints that will be considered as a reference for the labeling of the uncertain area. Moreover, these permanently labeled areas stabilize the algorithm by limiting the spread out of the appraisal errors. There are two main parameters: the inner and outer distance of the belt from the bounding box border. The outer limit embedding the object could be set up manually (we choose 4 pixels). However, the inner border will vary in relation to the object shape and the amount of outliers in its neighborhood.

3. LABELING ALGORITHM DESCRIPTION

We have thus taken advantage of these features to create the following keypoint labeling method. Be, for every given point p, $P_C(p)$, $P_M(p)$, $P_P(p)$, et $P_A(p)$ the probability of the point to belong to the object, respectively based on color, motion, position, and its associated point,. Be $L(p)$ the global likelihood that p belongs to the object. Be O_1 and O_2 the point areas deemed belonging respectively to the scenery and the object, and O_3 the belt of uncertain keypoints. Our algorithm initialize at the first frame the likelihood $L(p)$ according to their position $P_P(p)$ solely. For the other frames, if *p belongs to* O_1 *or* O_2, then $L(p)$ is the mean of the $P_P(p)$ and $P_A(p)$ probabilities. Else, if p is matched, its likelihood is calculated by averaging $P_C(p)$, $P_M(p)$, and $P_A(p)$. In the remaining case, $L(p)$ is not computed. An example is shown on Figure 2.

This method is quite efficient when the environment is cluttered, but this is not always the case. For instance, we could track a football player running on a uniform green lawn, where almost all the detected keypoints will then belong to the object. In this case, a basic algorithm considering the points inside (respectively outside) of the bounding box as object (respectively scenery) would be more efficient.

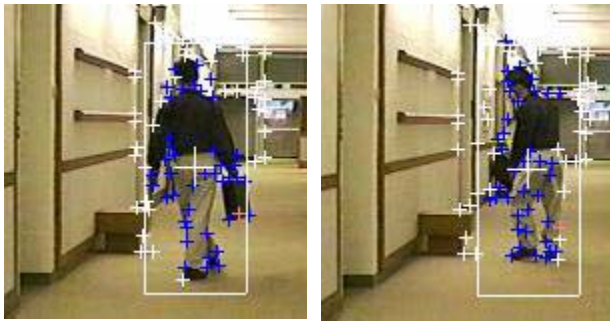

Figure 2: Labeling for frames 30 and 60 of the "surveillance" sequence. The "object" keypoints are in blue, those of the "background" in white, undetermined keypoints are in red.

The choice of which algorithm to use totally depends on a prior estimation of the *outlier* rate. The *outlier* rate *OutRate* is computed from the number *nbBck* of matched keypoints with a "background" labeled antecedent and the total number *nbKpts* of matched keypoints. Two factors are likely to perturb the accuracy of this computation: a too small number of keypoints and a large temporal instability (the trend, detailed in [7], of the keypoints to appear and disappear over time). An estimation using several previous frames (the importance of each frame decreasing with time) allows us to counterbalance the keypoint temporal instability. When the number of keypoints is too small, the *OutRate* is simply not computed. For each frame having a sufficient number (more than 10) of keypoints, the *OutRate* value is calculated as the mean of its current value plus the *nbBck/nbKpts* ratio. If this value is greater than the quantity of tolerated *outliers* T, then our labeling algorithm is executed. Else the basic labeling algorithm is performed. The T variable is directly dependent on the analyzed video. This test is done for each frame thereby detecting the passage from uniform to cluttered areas.

The major drawback of this algorithm comes from the definition of an "object" labeled central area in the bounding box. Indeed, this process is harmful for the occlusion detection. Nevertheless, our tracking algorithm, in order to be generic, relies on the necessary hypothesis of highly deformable object treatment. In consequence, if partial occlusions are efficiently handled, total occlusions are assimilated to an object change (with or without keypoint labeling) and have to be detected with an independent mechanism, thereby minimizing this flaw.

4. EXPERIMENTS

This section explores the labeling efficiency and the influence of the main parameters. The experiments have been conducted on color Harris-Laplace keypoints [10,11], but can be applied to any other type of keypoints.

The efficiency of the presented technique depends on two parameters. The first one is the size S of the analyzed area O_3 in relation with the bounding box, modeling the object compactness. A book perfectly fitting the border of the bounding box will need a value of *S* close to 0%. On the contrary, tracking a highly deformable object, like a person running, will produce a large label incertitude area. An optimal *S* parameterization should be a function of the object compactness. However, we are in the case of a generic tracking, and we do not have any knowledge of the object, neither any mean to assess to its compactness. Thus, we have experimentally determined an optimal size *S* of the analyzed belt without any a priori knowledge about the object. It is measured in relation to the biggest bounding box dimension (height or width). The benchmark is constituted of 6 video sequences with a cluttered background. Regarding the results presented in Figure 3, we have chosen an analyzed belt size equal to 20% of the bounding box.

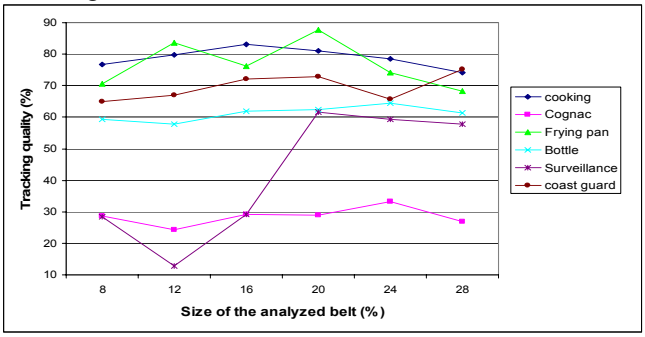

Figure 3: Influence of the size *S* analyzed belt size O_3 on the tracking quality.

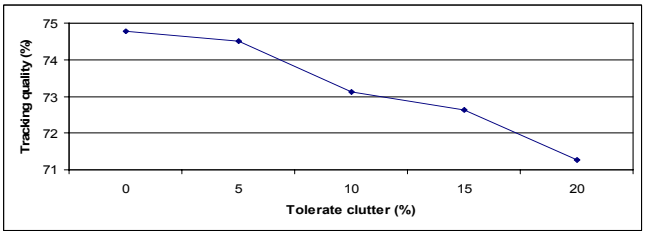

Figure 4: Influence of the quantity *T* of tolerated outliers on the tracking quality. Mean of the tracking quality calculated on 10 video sequences.

The second determinant variable for the algorithm behavior is the quantity *T* of *outlier tolerance*. This threshold marks the assumed limit of the object evolution between uniform and cluttered environment. It triggers in consequence the application of the appropriate labeling algorithm. In order to find the optimal *T* value, we have performed tracking quality tests on 10 cluttered and non cluttered video sequences for various values of this threshold. Figure 4 summarizes the results.

We can notice that 0% *outlier tolerance* means a permanent application of our labeling algorithm, whereas a tolerance superior to the outlier rate will be equivalent to the elementary labeling described in the introduction and Section 3. The results of the basic algorithm almost correspond to the case of the 20 % outlier tolerance of the Figure 4. Hence, this experience significantly validates our algorithm because the tracking quality increase is proportional to the application frequency of our labeling algorithm. In practice, we have adopted a 5% outlier tolerance, an inferior threshold implying useless computation for a meaningless 0.2% results growth.

5. CONCLUSION

This article has studied the differentiation between the object and the background in a keypoint based tracking application, where the object is defined by a bounding box. We proposed a keypoint labeling algorithm combining four features: label of the matched keypoints, color, motion, and position. The method has proved to significantly improve the accuracy of the tracking, on a set of diverse video sequences.

Lots of further improvements can be foreseen based on this first approach on the subject. For example, the bounding box repositioning can be enhanced based on the labeling. Also, the clutter estimation could be used to increase the results of the matching algorithms requiring such an input

6. REFERENCES

[1] Cavallaro A, Steiger O, Ebrahimi T, Tracking video objects in cluttered background, TCSVT, 15(4), pp.575-584, April 2005.

[2] Techmer A., Contour-based motion estimation and object tracking for real-time applications, In International Conf. on Image Processing, volume 3, pp. 648-651, Thessaloniki, Greece, 2001.

[3] Valette S, Magnin I, Prost R, Mesh-based video objects tracking combining motion and luminance discontinuities criteria, *Signal Processing archive*, *Vol 84(7) , pp. 1213-1224, July 2004.*

[4] Zhengyou Z. Parameter Estimation Techniques: A Tutorial with Application to Conic Fitting, *RR-267, Projet ROBOTVIS*, Sophia Antipolis, Octobre 1995.

[5] Charles V. Stewart, Robust Parameter Estimation in Computer Vision**,** SIAM review, 1999.

[6] R. Trichet and B. Mérialdo, "Generic Object Tracking for Fast Video Annotation", *VISAPP*, Barcelona, Spain, 2007.

[7] R. Trichet and B. Mérialdo, "Probabilistic Matching Algorithm for Keypoint Based Object Tracking Using a Delaunay Triangulation", *WIAMIS*, Santorini, Greece, 2007.

[8] Y. Bar-Shalom and T.E.FortMann, "*Tracking and Data Association*". New-York, Academic Press, 1988.

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

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

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