FCV2016, the 22nd Korea-Japan Joint Workshop on Frontiers of Computer Vision

Object Tracking based on Invariant Feature Points with Attributes

Takayoshi Yamashita
Chubu University
Kasugai , AICHI
Email: yamashita@cs.chubu.ac.jp

Yuji Yamauchi
Chubu University
Kasugai , AICHI
Email: yuu@vision.cs.chubu.ac.jp

Hironobu Fujiyoshi
Chubu University
Kasugai , AICHI
Email: hf@cs.chubu.ac.jp

***Abstract*—We will present in this paper a method based on invariant feature points with two attributes for robust object tracking against various appearance changes and occlusion. The first attribute, referred to as the affiliation attribute, determines whether each feature point belongs to the tracked object target region or to the background region. The use of this affiliation attribute on each feature point helped us eliminate the feature points in the target region that are affiliated to the background region. It also helped us determine whether occlusion had occurred or not. The second attribute, referred to as the memory attribute, determines the memory term of a feature point through its appearing frequency. The use of this memory attribute helped us minimize the model degradation of long-term stored feature points due to various appearance changes. Through the experiments we performed, we will show that these two attributes can help prevent tracking failures during appearance changes and occlusion where conventional methods are unsuccessful.**

I. Introduction

One of the main challenges of robust object tracking is the variation in appearance changes of the tracking target over time. These variations are due to changes in the shape of the object, changes in illumination, and occlusions. We think that focusing research in designing a flexible model, or adapt an existing model to appearance changes, is important. There are two approaches to designing a model for a method of object tracking: 1) treat the whole object as the target[2][3][4][5] 2) treat a collection of feature points as the target[9][11][13]. Furthermore, in order to track a target over time, a model needs to be sequentially updated based on an online learning method[6][7][8][10][12]. Although conventional methods have had success with tracking objects with changes in appearance, there are still challenges in real environment tracking. Firstly, conventional methods offer no robustness against occlusion. Tracking will often fail for occluded targets because there is not enough information to estimate their position. Secondly, traditional methods have difficulties in adapting to the changes in appearance of the target. As a result, such models are hindered with the discarded background information of the misaligned position of feature points resulting from the various appearance changes.

In this paper, we address these problems by presenting a method using invariant feature points with an affiliation attribute and a memory attribute. Our model consists of a set of invariant features points and is both a target model and a background model. Each feature point will have an affiliation and a memory attribute based on spatial and temporal relations. The *affiliation attribute* determines whether each feature point belongs to the target region or to the background region. The use of this affiliation attribute on each feature point helped us eliminate feature points in the target region that are affiliated to the background region. This affiliation feature also helped us determine whether occlusion had occurred or not. The second attribute, the *memory attribute*, decides the memory term of a feature point through its frequency of appearance. When we memorize something, recurrent information is stored in our long-term memory. In contrast, information that is only needed temporarily is stored in our short-term memory. Our method is based on this phenomena in order to use a memory attribute that determines the storing duration of each feature point. The method deals with the issues of occlusion and appearance changes of the target during tracking by using those two attributes with each feature point.

II. Tracking Model with Attributes

In conventional methods[11][13], the target position is estimated by matching pairs of feature points to a feature vector and their relative position to the center of the target. Grabner[13] proposed a method to estimate the target position from feature points in the background to deal with occlusion. However, tracking in a real environment is difficult because this method determines whether occlusion has occurred or not based on the number of feature points in the target region. As such, this method is not always successful in a situation where the target is occluded by an object that has many feature points. To address this issue, we apply the two attributes A_s and A_m on each feature point.

Firstly, A_s is defined as one of three affiliation attributes representing spatial relationship. A feature point in the target region is given a Target Feature (TF) attribute, and otherwise a Background Feature (BF) attribute. Furthermore, the feature points that hinder tracking in the target region is given a Discard Feature (DF) attribute. In concrete terms, the feature points that are affiliated with the background region are the ones causing the occlusion. Applying the DF attribute allows to identify and eliminate unneeded or useless feature points from the target region, and therefore reduce the tracking failures cause by model degradation during the updating process. The DF attribute helps understand the situation in the

surroundings and determine whether occlusion have occurred or not.

Secondly, A_m is defined as one of three memory attributes. These memory attributes, based on Squire's classification of memory[1], are the Iconic Memory (IM) attribute, the Short-term Memory (SM) attribute and the Long-term Memory (LM) attribute. We usually store iconic memory for about a second and short-term memory for about 20 seconds before forgetting it. However, short-term memory can be changed to long-term memory through a repeating process called*maintenance rehearsal*. Here, we apply the notion of imprinting memory (FM), which is a characteristic among birds, to the feature points in a frame and keep the feature points stored. The IM attribute is applied to the feature points in the current frame. Each feature point is stored for a particular duration depending on its attribute type. The memory attributes helps improve the tracking accuracy during various appearance changes.

We define the feature point attribute in the current frame as $F_{t,i} = \{x_i, \boldsymbol{v}_i, A_{s,i}, A_{m,i}\}$ and the stored feature points as $F_{m,j} = \{x_j, \boldsymbol{v}_j, \bar{x_j}, A_{s,j}, A_{m,j}\}$. The target position is estimated from the feature points with the TF and BF attributes. The feature points with the DF attribute are used to determine whether the target is occluded or not. Furthermore, the memory attributes are used based on the termination of the storing period of the feature points. This allows to avoid degradation of the model during online updating. We define object tracking based on feature points with attributes in Eq.(1):

$$\begin{aligned} p(x') =& \alpha \sum_{F_{t,i} \in FS_o} w(A_{m,i}) p(x'|x_{c,i}) \rho(F_{t,i}) \\ &+ (1-\alpha) \sum_{F_{t,i} \in FS_b} w(A_{m,i}) p(x'|x_{c,i}) \rho(F_{t,i}) \end{aligned} \tag{1}$$

A set of feature point FS is $FS = FS_o \cup FS_b \cup FS_d$. Each subset of feature points with TF, BF and DF attributes are defined respectively as FS_o, FS_b and FS_b. The target position is estimated with the subsets of feature points with the TF and BF attributes. α is the weight of each affiliation attribute. It decides whether to focus on the estimated position of the target or the background. $w(A_{m,i})$ is the weight of each memory attribute. It decides the importance of the attributes of a feature point in the current frame. Each weight is updated sequentially.

III. Framework of proposed method

The proposed framework is shown in Fig. 1. Our method matches the feature points extracted from the current frame $F_{t,i}$ to the one stored $F_{m,j}$. Each feature point receives its affiliation attribute and memory attribute from the matched feature point. The target position x_t^* is then estimated from the feature points with attributes. The feature vector, position and attributes of the stored feature points are updated through the matching feature points in the current frame. Unneeded or useless feature points are removed from the stored feature points FP depending on the memory duration of each attribute. New matching feature points are stored in FP.

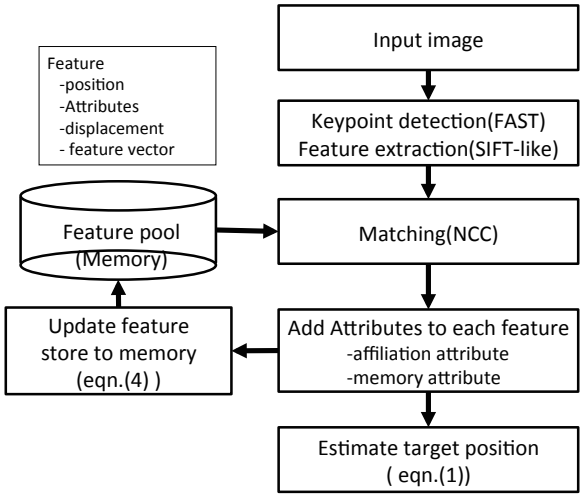

Fig. 1. Tracking flow of proposed method.

A. Feature Extraction

The feature points are extracted with a FAST detector [14] and the descriptors are based on similar SIFT[13]. The gradient histogram with 16 bins is calculated from a 30×30 region. The region is rotated in the direction that had the maximum value in the histogram. For the descriptor, the region is divided into 3×3 blocks and the gradient histogram is calculated with 8 bins for each block. The descriptor is the feature vector $\boldsymbol{v}$ with 72 dimensions and is linear with the histograms of all blocks. The feature vectors $\boldsymbol{v}_i$ $\boldsymbol{v}_j$ are stored in the memory recursively. The feature point that is matching scores by normalized cross-correlation larger than the threshold TH_m are stored in FS.

B. Understanding surrounding with the affiliation attribute

The number of feature points with the TF attribute is diminished when the target is occluded. It is difficult to estimate the target position with a reduced set of feature points. We employ the weight that determines the attention whether estimated position of the target or the background based on the number of feature points with DF and TF attributes. The weight α, is defined as the value of understanding the surroundings, is as follows:

$$\alpha = \begin{cases} 1 - \frac{(|FS_d|)^2}{(|FS_o|)^2+1} & |FS_o| \geq |FS_d| \\ 0 & otherwise \end{cases} \tag{2}$$

The weight attention to estimated position of the background when the number of feature points with the DF attribute is an increase. On the other hand, the weight attention to the estimated position of the target when the number of feature points with the TF attribute is an increase

C. Importance of the Memory Attribute

The weight of the memory attributes rises according to the number of matching pairs. The weight of each memory

attribute is defined in Eq.(3).

$$w(A_{m,i}) = \begin{cases} \frac{|FS_I|}{|FS|} & (A_{m,i} = IM) \\ \frac{|FS_S|}{|FS|} & (A_{m,i} = SM) \\ \frac{|FS_L|}{|FS|} & (A_{m,i} = LM) \\ \frac{|FS_F|}{|FS|} & (A_{m,i} = FM) \end{cases} \quad (3)$$

The subsets for the feature points with the IM attribute, the SM attribute, the LM attribute and the FM attribute are respectively FS_I, FS_S, FS_L and FS_F.

D. Target Position Estimation

The target position is determined through a maximum combination of the position likelihood $p(x'|x_{c,i})$ of each feature point. $p(x'|x_{c,i})$ is assumed with a Gaussian distribution with estimated target position $x_{c,i}$. The likelihood of the target position $p(x')$ is calculated from the estimated position $x_{c,i}$ and $p(x'|x_{c,i})$ as shown in Eq.(1).

E. Update of Stored Feature Points

The stored feature points are updated according to the matching feature points. The feature points include in the target region receive the TF attribute. If the feature points in the target region match feature points with the BF attribute, the DF attribute is given. If the feature points $F_{t,i}$ match feature points $F_{m,j}$ with the IM attribute, the SM attribute is given. The feature point is given the LM attribute instead of the SM attribute when the feature point is matched more than T_m times. The feature points $F_{t,i} \notin FS$ that do not match with any feature point is given the IM attribute. The position and the feature vector of the stored feature points $F_{m,j}$ are respectively updated with matching $F_{t,i}$ as $y_j = x_i$ and $\boldsymbol{v}_j = \boldsymbol{v}_i$. The displacement is updated with the previous $\bar{x}_j$, current $(x^* - x_i)$ and coefficient β as defined in Eq.(4).

$$\bar{x}_j \quad = \quad \beta(x^* - x_i) + (1 - \beta)\bar{x}_j \quad (4)$$

The feature points with the IM attribute that do not match any feature points for more than T_i are removed from the stored memory FS. The feature points with the SM attribute and the LM attribute are also removed if they do not match any feature point for more than T_s and T_l respectively. The number of matching points is cumulated even intermittently. On the other hand, the number of mismatching points is reset when a feature point can be matched to $F_{m,j}$.

IV. Experiments

A. Experiment protocol

We compare our method with feature point based methods [11][13] for a common dataset [12] and our dataset that includes six home videos. The common dataset [12] includes unnatural motions, like quick direction changes, etc. We therefore add a dataset that includes natural appearance change and occlusion found in home videos. The settings for the comparison method are based on the original literature.

TABLE I
Error of center position in public DB[pixel].

Video Clip	ST[11]	SP[13]	Ours
David Indoor	44	32	21
Sylvester	18	21	12
Occluded Face	36	29	18
Occluded Face2	45	58	18
Girl	39	38	24
Tiger	34	31	18
Tiger2	28	24	17
Coke Can	34	28	25
average	34	32	19

TABLE II
Error of center position in home video[pixel].

Video Clip	ST[11]	SP[13]	Ours
HumanPose1	34	30	15
HumanPose2	59	43	21
HumanPose3	23	24	11
DogPose	42	38	16
PartialOcclusion	30	21	9
FullOcclusion	45	35	17
average	38	31	14

[11] uses SURF while [13] uses Harris operator and SIFT-like descriptor. The parameters of our method are defined by a preliminary experiment. TH_m is 0.7, the coefficient β is 0.7, the number of matching T_m is 5, and the number of mismatches of IM, SM and LM are respectively 5, 30 and 100. We measure the average error of the center position between the estimated position and the ground truth in the whole sequence.

B. Experiment Results with Common Dataset

Table 1 shows the results for the tracking errors with the common dataset. The results for the method with the best performance are indicated in red in each sequence. As shown, our method is achieved the best performance in five sequences and is second in the other three. It is also achieved the lowest error average in all sequences.

The tracking results are shown in the frames in Fig. 2. The blue rectangles indicate estimation results for feature points with the TF attribute while the yellow rectangles indicate the estimation results for features points with the BF attribute. Our method is successful even for changes in appearance due to face pose or illumination conditions, as shown in Fig. 2(a)(c). Our method also is successful in estimating the target position of the feature points with the BF attributes even during occlusion, as shown in Fig. 2(b). However, our method had difficulty in successfully tracking the target when it si moving quickly during occlusion, as shown in Fig. 2 (d). It is very difficult to estimate the target position from the feature points with the BF attribute.

C. Experiment Results with Home Videos

Table II shows the results for the tracking errors with the home videos. Our method is achieved the best performance

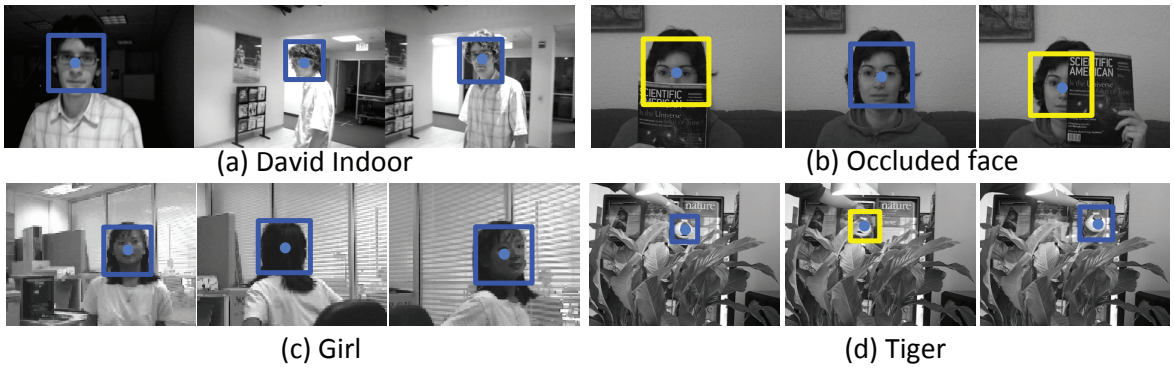

Fig. 2. Tracking result of proposed method in public DB.

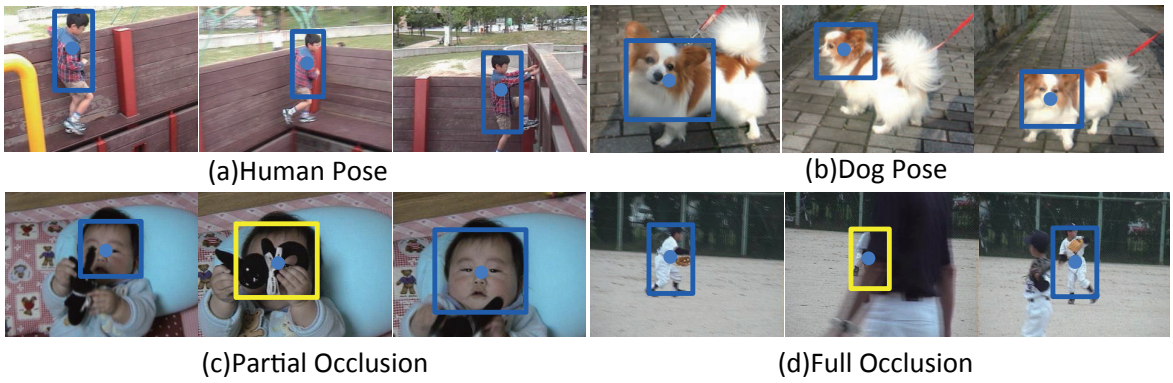

Fig. 3. Tracking result of proposed method in Home video.

with this dataset. The tracking results are shown in the frames in Table II. Our method is successful in tracking changes in posture of humans and dogs, as shown in Fig. II (a)(b). Our method is also successful in estimating the target position with feature points with the BF attribute while it is partially or fully occluded, as shown in Fig. II (c)(d).

V. DISCUSSION

We discuss the affiliation attribute change in the vicinity of where occlusion happens. Figure 4 shows the feature point in certain frames. While the yellow rectangle indicates that the number of feature points with DF attribute is larger than the feature points with TF attribute, the blue rectangle indicates the other cases. In Fig.4(b), DF attribute is given to feature point, since the feature points of the book are part of the background. In Fig.4(c), the target position is estimated by the feature points with BF attribute since the number of feature points with DF attribute is larger than the ones with TF attribute. After occlusion by the book in Fig.4(d)(e), BF attribute is given to the feature points and the estimation of the target position is done by the feature points with TF attribute.

Figure 5 shows the feature points with memory attribute in each frame. In Fig.5(a) and (d), many feature points with the FM attribute appear since the target is similar to the initial one. As shown in Fig.5(b) and (c), many IM and SM attributes kept short term are given to the feature points during target movement or very small shape variation. On the other hand, the LM and FM attributes that are kept long term are given to the feature points when appearance changes through face pose. From this analysis, the IM and SM attributes correspond to very small shape variations and the LM and FM attributes correspond to larger appearance changes. Furthermore, the memory attribute prevents model degradation due to model updating and tracks well even with appearance variations.

VI. CONCLUSION

In this paper, we present an object tracking method based on feature points with affiliation and memory attributes. Our method allows to be successful with robust object tracking against appearance changes and during occlusion by using the attributes of feature points to estimate the target position. In future works, we will apply these attributes through different approaches, e.g. tracking by detection, etc.

REFERENCES

[1] L.R. Squire, “Memory and Brain”, Oxford University Press, 1987.
[2] M. Isard and A. Blake, “Contour tracking by stochastic propagation of conditional density”, Proc. European Conf. on Computer Vision, pp.343–356, 1996.

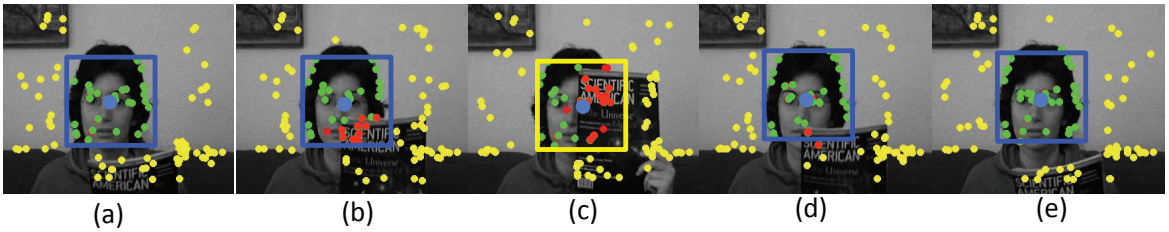

Fig. 4. The feature points of each affiliation attribute in occlusion scene. Green point is TF attribute, yellow point is TF attribute, red point is DF attribute and blue point is center position.

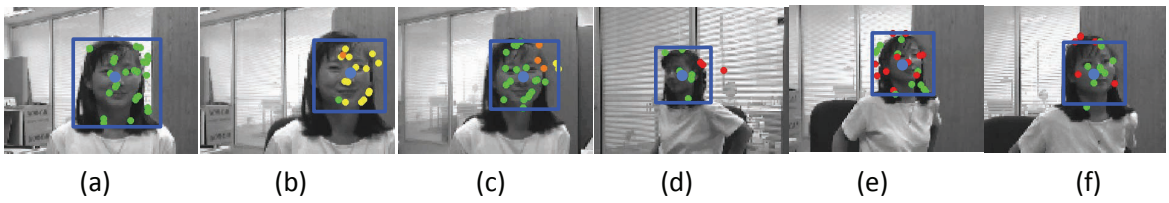

Fig. 5. The feature points of each memory attribute in pose change scene. A green point is the FM attribute, a yellow point is the IM attribute, a red point is the LM attribute and a blue point is the center position.

[3] M. J. Black and A. D. Jepson, “Eigentracking: Robust matching and tracking of articulated objects using a view-based representation”, Intl. Journal of Computer Vision, Vol.26, No.1, pp.63–84, 1998.
[4] S. Birchfield, “Elliptical head tracking using intensity gradients and color histograms”, Proc. IEEE Conf. on Computer Vision, pp.232–237, 1998.
[5] D. Comaniciu, V. Ramesh, P. Meer, “Kernel-based object tracking”, IEEE Trans. PAMI, Vol.25, No.5, pp.564–577, 2003.
[6] R. T. Collins, Y.Liu, “Online selection of discriminative tracking features”, IEEE Trans. PAMI, Vol.27, No10, pp.1631–1643, 2005.
[7] H. Grabnber and H. Bischof, “On-line Boosting and Vision”, Proc. IEEE Conf. on Computer Vision, pp.260–267, 2006.
[8] S. Avidan, “Ensemble Tracking”, IEEE Trans. PAMI, Vol. 29. No.2, pp.261–271, 2007.
[9] M. Yang, J. Yuan, Y. Wu, “Spatial Selection for Attentional Visual Tracking”, Proc. IEEE Conf. on Computer Vision, pp.1–8, 2007.
[10] D. A. Ross, J. Lim, R-S. Lin, M-H. Yang, “Incremental Learning for Robust Visual Tracking”, Intl. Journal of Computer Vision, Vol.77 No.3, pp.125–141, 2008.
[11] W. He, T. Yamashita, H. Lu, S. Lao, “SURF Tracking”, Proc. IEEE Intl. Conf. on Computer Vision, pp.1586-1592, 2009.
[12] B. Babenko, H-M Yang, S. Belongie, “Robust Object Tracking with Online Multiple Instance Learning”, IEEE Trans. PAMI, Vol.33, No.8, pp.983-990, 2010.
[13] H. Grabner, J. Matas, L. Van Gool, “Tracking the invisible: Learning where the object might be”, Proc. IEEE Conf. on Computer Vision, pp.1–8, 2010.
[14] E. Rosten, R. Porter, T. Drummond, “Faster and better: a machine learning approach to corner detection”, IEEE Trans. PAMI, Vol.32, No.1, pp.105–119, 2010.