EDITED FILM ALIGNMENT VIA SELECTIVE HOUGH TRANSFORM AND ACCURATE TEMPLATE MATCHING

Xiaomeng Wu, Takahito Kawanishi, Minoru Mori, Kaoru Hiramatsu, and Kunio Kashino

NTT Communication Science Laboratories, NTT Corporation

ABSTRACT

Edited film alignment is the post-production process of finding small parts of unedited footage that temporally and spatially match an edited film. The huge amount of data to be processed makes significant downsampling of the videos essential in real-life applications. Simultaneously, professional users demand that the task be achieved with frame and pixel-level accuracy. We propose a novel selective Hough transform (SHT) and an accurate template matching method to address the difficult trade-off between accuracy and scalability. For robust temporal alignment, SHT investigates the selectivity of frame-level similarities and advantageously reduces the weights of mismatches. The template matching method encompasses spatial Hough transform and sum of squared differences (SSD) minimization. SSD is efficiently approximated by exploiting the second-order derivative of image intensity. Experiments conducted on real-world data show the superiority of our methods.

Index Terms — Video copy detection, Hough transform, template matching, sum of squared differences (SSD), approximate SSD

1. INTRODUCTION

Given a collection of unedited footage and an edited film, edited film alignment is the process of determining, for each frame of the edited film, the take, the frame, and the spatial position at which the frame concerned occurs in the collection. There is a growing need for an automated edited film alignment system, and professional users demand that the task be achieved with frame and pixel-level accuracy. The system should also be able to address reframing, e.g., pan, tilt, and zoom. Figure 1 shows a common film editing framework. The pieces of raw, unedited footage recorded during the making of a motion picture are called dailies. They consist of multiple takes. A take refers to each filmed version of a particular scene. The film editor works with the dailies, selecting shots from takes and combining them into a sequence to create a finished motion picture.

Edited film alignment relates to two research agendas known as video copy detection and template matching. Video copy detection methods arrange a query and each reference video to identify the video and the temporal position at which the query occurs. Douze et al. [1] conducted pairwise frame matching and showed that the temporal offset of the query can be found with a 1DOF Hough transform. Although global features showed great efficiency as regards frame matching without a viewpoint change [2, 3], more researchers have paid attention to local features [4, 5] or local feature geometry [1, 6] when addressing geometric transformations, e.g., caused by reframing. However, the huge amount of data to be processed makes significant downsampling of frames essential in real-life applications, which degrades the accuracy due to unavoidable mismatches. This is one of the main issues we address in this study.

We particularly thank Mr. David Fincher, who gave us permission to use all the materials required for this research. Special thanks go to the staff of PIX System for their generous support.

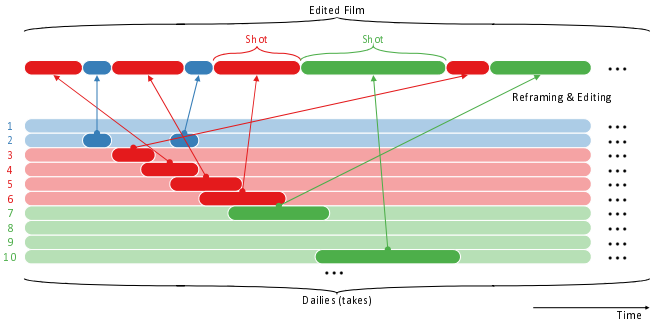


Fig. 1. Film editing. Each color indicates a single scene.

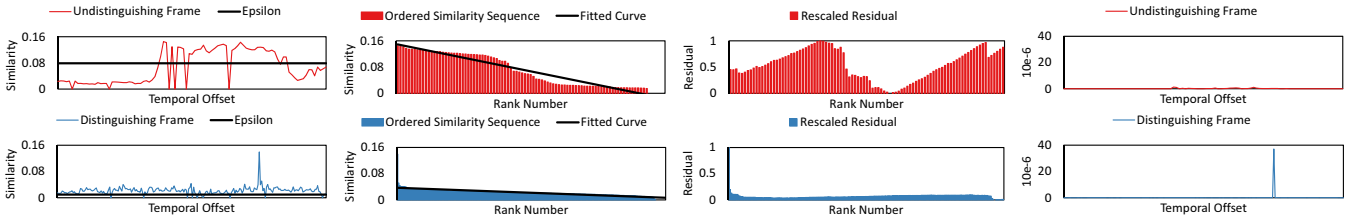
Template matching methods find small parts of an image that match a template image. The simplest way is to consider a huge number of geometric transformations and to find the transformation that minimizes the SSD error of pixels. Previous efforts were put into reducing either the transformation candidates [7–9] or the reference pixels [10]. Lowe [11] showed that a spatial Hough transform based on local features, followed by a least-squares solution, enables more efficient template matching. However, the least-squares solution tends to deliver a suboptimal approximation of the true transformation when the errors in feature correspondences are non-negligible. This leads to another issue that we focus on in this study.

We propose a novel edited film alignment approach to address the above issues that involves a selective Hough transform (SHT) and an accurate template matching method. Aiming at robust temporal alignment, the SHT investigates the selectivity of frame-level similarities and advantageously reduces the weights of mismatches (Section 2). The template matching method encompasses a spatial Hough transform and SSD minimization to achieve spatial alignment with pixel-level accuracy (Section 3). To accommodate the increase in complexity, for each transformation candidate we approximate the SSD error using a sublinear algorithm that examines only a small number of pixels with strong responses in terms of their second-order derivatives. In Section 4, we report our experiments conducted on real-word data provided by a film production company. Future directions are discussed in Section 5.

2. EDITED FILM ALIGNMENT IN TEMPORAL DOMAIN

2.1. Overview

We assume that the shot boundaries of the edited film are detected in advance. Given a shot, the problem is to determine the take T^* and the temporal offset Δt^* at which the shot was selected. If the offset is



(a) Similarity sequence (b) Ordered similarity sequence (c) Rescaled residual sequence (d) Selectivity function output

Fig. 2. Examples of similarity sequences, ordered similarity sequences, rescaled residual sequences, and selectivity function output.

constant over time, we can solve this problem with a 2DOF Hough transform. Let I_i and t_i denote each shot frame and its time stamp, respectively. For each take frame I'_j, we are given the take ID T_j and time stamp t'_j. The temporal transformation derived from each pair (I_i, I'_j) form an ordered pair $(T_j, \Delta t_{i,j})$ with $\Delta t_{i,j} = t'_j - t_i$. The two parameters in this ordered pair span a 2DOF voting map. The Hough transform is realized by voting all (I_i, I'_j) into the voting map according to their temporal transformations. The vote is defined as the similarity between I_i and I'_j and is denoted by $s_{i,j}$. Take T^* can thus be determined by $T^* = \arg\max_T(\max_{\Delta t} h(T, \Delta t))$, and the offset is determined by $\Delta t^* = \max_{\Delta t} h(T^*, \Delta t)$. Here, $h(\cdot,\cdot)$ is the total vote of each bin in the voting map. Because of its great efficiency, we adopt a bag-of-visual-words (BOVW) model [12, 13] to define the pairwise frame similarity $s_{i,j}$.

2.2. Selective Hough Transform

The huge amount of data to be processed makes significant downsampling of dailies vital in real life applications because of the efficiency requirement. This downsampling weakens the temporal consistency constraint imposed on offset detection and degrades the accuracy due to unavoidable mismatches.

Given a shot frame I_i and its corresponding take T^* with n frames, we look at the similarities between I_i and all take frames I'_j:

$$\mathbf{s}_i = (s_{i,1}, s_{i,2}, \cdots, s_{i,j}, \cdots, s_{i,n}). \tag{1}$$

We call $\mathbf{s}_i$ the similarity sequence of I_i. Two examples of $\mathbf{s}_i$ are shown in Fig. 2a. The red curve shows the similarity sequence of an ambiguous frame, which votes for multiple hypotheses with roughly equivalent voting values. In contrast, the blue curve corresponds to a more distinguishing frame and is expected to be more useful for offset detection. For robust offset detection, it is necessary to reduce the weights of ambiguous frames, and to accentuate the weights of distinguishing frames. Douze et al. [1] focused on a similar issue and proposed weighting the similarity on the basis of the l_1-norm or the maximum frame-level similarity. However, the direct application of this method may overemphasize the similarity between I_i and the frames of irrelevant takes, resulting in a suboptimal balance between take detection and temporal alignment.

Here, we consider a thresholded, polynomial selectivity function $\sigma_\alpha : \mathbb{R} \to \mathbb{R}^+$ of the form

$$\sigma_\alpha(s_{i,j}) = \begin{cases} |s_{i,j} - \varepsilon_i|^\alpha & \text{if } s_{i,j} > \varepsilon_i \\ 0 & \text{else} \end{cases} \tag{2}$$

where $\varepsilon_i \in [0, s^{(i)}_{\max}]$ with $s^{(i)}_{\max}$ being the maximum of $\mathbf{s}_i$ and $\alpha > 0$. The SHT replaces each $s_{i,j}$ with $\sigma_\alpha(s_{i,j})$ and casts the latter for the corresponding $(T_j, \Delta t_{i,j})$. A larger α increases the selectivity and implicitly yet greatly reduces the weights of potential mismatches. We want to adaptively select ε_i such that a larger ε_i, which minimizes the contribution of votes, is used for ambiguous frames, and vice versa. The similarity sequence of an ambiguous frame is usually shaped like a uniform distribution, while that of a distinguishing frame is much more single-peaked. This motivates us to define

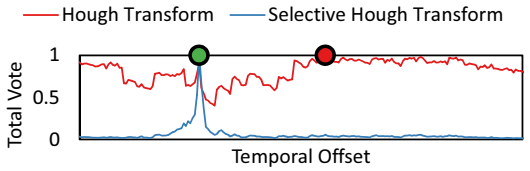


Fig. 3. Example voting maps. The red and green circles are the offset detected by the Hough transform and the true offset, respectively.

$$\varepsilon_i = w_i \times s^{(i)}_{\max} \tag{3}$$

and to select $w_i \in [0, 1]$ by exploiting the non-uniformity of $\mathbf{s}_i$.

In particular, we sort all $s_{i,j} \in \mathbf{s}_i$ in descending order, as shown in Fig. 2b. We then construct the first degree polynomial equation of the ordered similarity sequence. In the next step, the fitted curve, which is expected to approximate the *distribution* of off-peak similarities that are far from the head of $\mathbf{s}_i$, is subtracted from the ordered similarity sequence. This is to eliminate the effect of off-peak similarities on the selection of w_i and to increase the margin of the non-uniformity of $\mathbf{s}_i$ between ambiguous and distinguishing frames. From Fig. 2c, we can see that the rescaled residual sequence of the ambiguous frame has a much larger area under the curve (AUC) than that of the distinguishing frame. We thus define w_i by

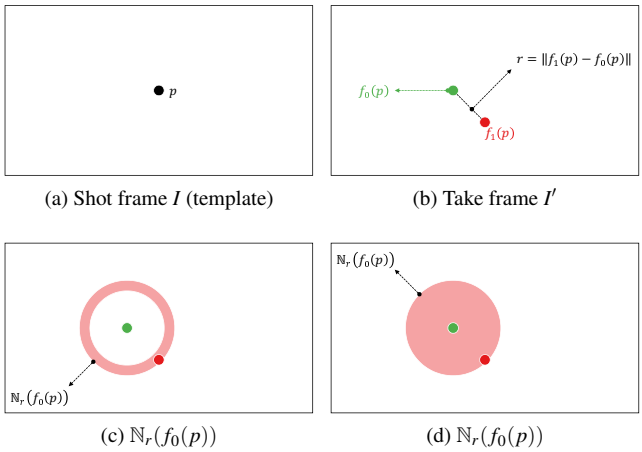
$$w_i = \frac{\mathrm{AUC}(\bar{\mathbf{r}}_i)}{n} \tag{4}$$

where $\bar{\mathbf{r}}_i$ denotes the rescaled residual sequence. The weight is substituted into Eq. 3 to determine the adaptive threshold ε_i, as shown by the heavy black lines in Fig. 2a. The output of the selectivity function is shown in Fig. 2d. We can see how the SHT suppresses the function of the ambiguous frame. Figure 3 compares two voting maps obtained with the Hough transform and the SHT, where the SHT allows a much easier prediction of the temporal offset.

3. ACCURATE TEMPLATE MATCHING

3.1. Spatial Hough Transform and Revalidation

Once the temporal offset has been detected, Lowe's method [11] can be used to detect the geometric transformation between each shot



(a) Shot frame I (template) (b) Take frame I'

(c) $\mathbb{N}_r(f_0(p))$ (d) $\mathbb{N}_r(f_0(p))$

Fig. 4. Investigation of usefulness of pixels for template matching.

frame and its counterpart in the corresponding take. Given a template I (shot) and an image I' (take), the method outputs a set of confident feature correspondences. Lowe proposed applying least squares to the coordinates of these correspondences to accurately estimate the geometric transformation.

To avoid suboptimal approximation, we generate a transformation candidate from each confident correspondence [11] and revalidate all the candidates to find a transformation that comes close to minimizing the SSD between I and I'. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ denote the transformation function that maps a coordinate $p \in I$ to its counterpart $f(p) \in I'$. Let P be a subset of pixels in I. The SSD between I and I' is of the form

$$\mathrm{SSD}_f(I, I') = \frac{1}{|P|} \sum_{p \in P} |I'(f(p)) - I(p)|^2. \tag{5}$$

Instead of using all the pixels, we approximate SSD by inspecting a small fraction, e.g., 0.005% (100 samples for FHD resolution), of pixels from I. In Section 3.2, we investigate the usefulness of pixels for minimizing the negative effect of the approximation.

3.2. Second-Order Derivative of Image Intensity

Given a pixel $p \in I$ (Fig. 4a), let f_0 and f_1 be the true transformation and an incorrect transformation, respectively. The corresponding coordinates $f_0(p)$ and $f_1(p)$ in I' are shown in Fig. 4b with green and red dots. The transformations f_0 and f_1 are distinguishable with $I(p)$ only if

$$|I'(f_1(p)) - I(p)|^2 > |I'(f_0(p)) - I(p)|^2. \tag{6}$$

Assuming that $I(p) \approx I'(f_0(p))$, we can obtain the necessary condition of Eq. 6:

$$|I'(f_1(p)) - I'(f_0(p))| > 0. \tag{7}$$

The left part of Eq. 7 can be understood as a cue that indicates how useful $I'(f_0(p))$ is for rejecting f_1. Let $u(f_0(p))$ denote this cue. Let r be the distance between $f_0(p)$ and $f_1(p)$, and $\mathbb{N}_r(f_0(p))$ be the set of coordinates (Fig. 4d) that are at most r pixels away from $f_0(p)$. If we broaden our scope from $f_1(p)$ to all the coordinates in $\mathbb{N}_r(f_0(p))$, we can extend $u(f_0(p))$ by

$$u(f_0(p)) = \sum_{q \in \mathbb{N}_r} |I'(q) - I'(f_0(p))| \tag{8}$$

$$\geq \left| \sum_{q \in \mathbb{N}_r} I'(q) - I'(f_0(p)) \right| \tag{9}$$

where Eq. 9 is the lower bound of Eq. 8. To maximize the accuracy of the approximation, we wish to find the pixels with the largest response in terms of the right hand side of Eq. 9.

Note that the computation of Eq. 9 for all the pixels in I' equals the convolution of I' with a Laplacian operator whose size depends on r. The response of Eq. 9 of a given pixel thus corresponds to its second-order derivative of intensity. This important connection inspired us to utilize commonly used second-order approaches, e.g., LoG [14] and DoG [11], for effective pixel sampling.

Table 1. Dataset.

Dataset	Resolution	#shot	#take	#frame (shot)	#frame (take)
HOC	1920×1080	25	110	2,088	216,465

3.3. Pixel Sampling

We adopt a Hessian affine region detector [15, 16] for this purpose, which corresponds to a hybrid operator between the Laplacian and the determinant of the Hessian operator. The scale of the points detected by this approach is related to the radius r in Fig. 4 and plays an important role as an alternative barometer of pixel usefulness. An interest point with a smaller scale marks a great distinction from its neighboring pixels, and so is advantageously more sensitive to the variation between similar transformations. Therefore, we sort all interest points in ascending order of their scales and select the top points for the SSD approximation. In Section 3.2, we discussed the usefulness of the pixels by using the take frame I' as an example. In practice, since the number of shots is much smaller than the number of takes, we sample pixels only from the shot for greater efficiency.

Note that our method uses the pixels that lie centrally in the interest points only for the SSD approximation. No feature descriptor or correspondence is required at this stage. Our method is different from conventional local feature-based methods, for example Lowe's method [11], which rely on corresponding feature points for computing the global transformation.

4. EXPERIMENTS

4.1. Implementation

Our methods are built on top of local feature-based image indexing. We used a rotation-variant Hessian affine region detector [15, 16] to extract local features, and used root SIFT descriptors [17]. A visual vocabulary containing 1M visual words was trained on an independent dataset called Oxford Buildings [13]. An inverted index was used to efficiently compute the cosine similarity $s_{i,j}$ between BOVW histograms, and to find correspondences in Lowe's method [11].

4.2. SHT

We evaluated the SHT on a real-world dataset, called *House of Cards (HOC)*, provided by a film production company. Table 1 shows its statistics. Given a shot and the voting map obtained with a Hough

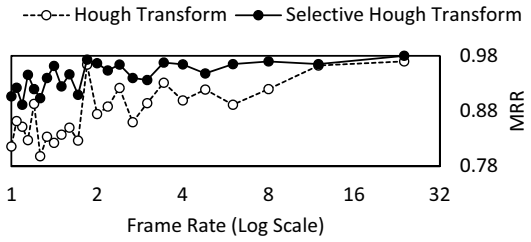


Fig. 5. MRR vs. frame rate (take). $\alpha = 5$.

transform, we sort all takes and all the temporal offsets in the relevant take according to their scores (Section 2.1) in the voting map. This gives us two ranking lists; one for take detection and one for temporal alignment. We use the mean reciprocal rank (MRR) [18] and the accuracy at one rank as the evaluation measure.

Figure 5 shows the relationship between MRR and the frame rate of takes. On the shot side, we used all the frames to achieve frame-level accuracy. Although the MRR of the SHT also degrades as we decrease the frame rate, it consistently outperforms the conventional Hough transform for all the tested frame rates.

Table 2. Average MRR and average accuracy over all tested frame rates (take). $\alpha = 5$.

Methods	Take Detection		Temporal Alignment	
	MRR	Accuracy	MRR	Accuracy
Hough Transform	**1**	**1**	.880	.842
Max Weighting [1]	**1**	**1**	.888	.847
l_1-Norm Weighting [1]	.810	.647	.926	.892
SHT (Otsu's Method [19])	**1**	**1**	.852	.803
SHT (MET [20])	**1**	**1**	.925	.887
SHT (Our Method)	**1**	**1**	**.944**	**.920**

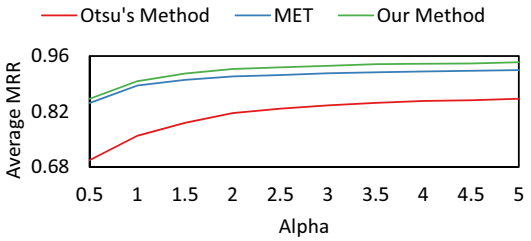


Fig. 6. Average MRR vs. α.

Table 2 compares SHTs with the two weighting methods used in Douze et al.'s method [1]. We also compared our thresholding method with Otsu's method [19] and maximum entropy thresholding (MET) [20], which are widely used in image processing. The l_1-norm weighting [1] performed reasonably well for temporal alignment, but degraded the accuracy of take detection due to the overemphasis of the similarity between shots and irrelevant takes. Otsu's method and MET underperformed our method because they only work on bimodal distributions, which is not the case with the frame-level similarity sequence focused on in an SHT. The relationship between MRR and the parameter α (Eq. 2) is shown in Fig. 6. The SHT treats all non-zero votes as one when $\alpha = 0$ and reduces to max pooling when $\alpha \to \infty$. We set $\alpha = 5$ for all previous evaluations.

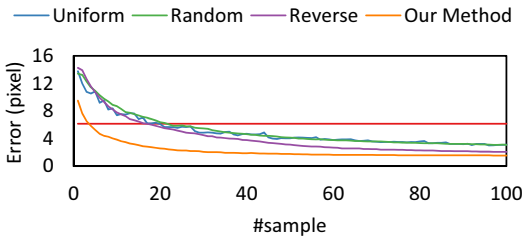


Fig. 7. Template matching error vs. #sample (HOC*). The red line indicates the error of Lowe's method [11].

4.3. Template Matching

Since it is difficult to manually label the geometric transformation between shot and take frames, we created a synthetic dataset based on HOC to evaluate template matching. For each shot frame, its counterpart in the relevant take is randomly projected with a similarity transformation. The resulting dataset contains 2,088 pairs of template and image frames, and is called HOC*. Given a frame pair, our method outputs the coordinates of the bounding box where the template occurs. We look at the maximum difference between the estimated and the true coordinates. The average of the maximums over all frame pairs is defined as the template matching error.

In Fig. 7, we compare our method with Lowe's method [11], uniform sampling, random sampling, and a reversed version of our method. Uniform sampling and random sampling correspond to the solutions adopted by Zhang and Akashi [9] and by Korman et al. [8], respectively. *Reverse* sampling sorts the interest points with their scales in descending rather than ascending order.

By inspecting only a small fraction of the pixels, e.g., 0.002% for 40 samples, all the methods that encompasses the spatial Hough transform and SSD minimization outperformed Lowe's method [11]. With a larger number of samples, both *reverse* sampling and our method outperform uniform and random sampling, which demonstrates the validity of sampling pixels on the basis of their second-order derivatives. However, *reverse* sampling did not match our method in terms of accuracy, especially when the number of samples was lower than 40. This supports our discussion in Section 3.3 on the radius r in Fig. 4. With 100 samples, we obtained a template matching error of 1.55 pixels and reduced the error of Lowe's method by 75%.

5. CONCLUSIONS AND FUTURE DIRECTIONS

We proposed a novel selective Hough transform (SHT) and a template matching method for edited film alignment. The SHT suppresses the negative impact of mismatched frames and allows the easier prediction of temporal offsets. We investigated the relationship between template matching and the second-order derivative of image intensity. We also showed how the second-order derivative and the scale of interest points can be used as barometers of pixel usefulness for effective pixel sampling. In the future, we plan to explore the reliability of our method by extending the experiments to larger-scale data. During template matching, we utilized a Hessian affine region detector [15, 16] for pixel sampling. It will be interesting to explore whether we can improve our approach by directly manipulating the responses of the Laplacian and the determinant of the Hessian operator. This will be the subject of a future study.

6. REFERENCES

[1] Matthijs Douze, Herve Jegou, and Cordelia Schmid, "An image-based approach to video copy detection with spatio-temporal post-filtering," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 257–266, 2010. 1, 2, 4

[2] Chenxia Wu, Jianke Zhu, and Jiemi Zhang, "A content-based video copy detection method with randomly projected binary features," in *CVPR Workshops*, 2012, pp. 21–26. 1

[3] Minoru Mori, Takayuki Kurozumi, Hidehisa Nagano, and Kunio Kashino, "Video content detection with single frame level accuracy using dynamic thresholding technique," in *ICPR*, 2014, pp. 2560–2565. 1

[4] Alexis Joly, Olivier Buisson, and Carl Frélicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 293–306, 2007. 1

[5] Chih-Yi Chiu, Tsung-Han Tsai, Yu-Cyuan Liou, Guei-Wun Han, and Hung-Shuo Chang, "Near-duplicate subsequence matching between the continuous stream and large video dataset," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1952–1962, 2014. 1

[6] Xiaoguang Gu, Dongming Zhang, Yongdong Zhang, Jintao Li, and Lei Zhang, "A video copy detection algorithm combining local feature's robustness and global feature's speed," in *ICASSP*, 2013, pp. 1508–1512. 1

[7] Bogdan Alexe, Viviana Petrescu, and Vittorio Ferrari, "Exploiting spatial overlap to efficiently compute appearance distances between image windows," in *NIPS*, 2011, pp. 2735–2743. 1

[8] Simon Korman, Daniel Reichman, Gilad Tsur, and Shai Avidan, "Fast-match: Fast affine template matching," in *CVPR*, 2013, pp. 2331–2338. 1, 4

[9] Chao Zhang and Takuya Akashi, "Fast affine template matching over Galois field," in *BMVC*, 2015, pp. 121.1–121.11. 1, 4

[10] M. Hashimoto, T. Fujiwara, H. Koshimizu, H. Okuda, and K. Sumi, "Extraction of unique pixels based on co-occurrence probability for high-speed template matching," in *International Symposium on Optomechatronic Technologies (ISOT)*, 2010, pp. 1–6. 1

[11] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 1, 2, 3, 4

[12] Josef Sivic and Andrew Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477. 2

[13] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007. 2, 3

[14] Tony Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998. 3

[15] Krystian Mikolajczyk and Cordelia Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004. 3, 4

[16] Michal Perdoch, Ondrej Chum, and Jiri Matas, "Efficient representation of local geometry for large scale object retrieval," in *CVPR*, 2009, pp. 9–16. 3, 4

[17] Relja Arandjelovic and Andrew Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918. 3

[18] Ellen M. Voorhees, "The TREC-8 question answering track report," in *TREC*, 1999. 4

[19] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. 4

[20] J. N. Kapur, Prasanna K. Sahoo, and Andrew K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985. 4