REFA3D: ROBUST SPATIO-TEMPORAL ANALYSIS OF VIDEO SEQUENCES

Manuel Grand-Brochier, Christophe Tilmant and Michel Dhome
Laboratoire des Sciences et Matriaux pour l'Electronique et d'Automatique (LASMEA), UMR 6602 UBP-CNRS, 24 Avenue des Landais, 63177 Aubière, France

Keywords: Ellipsoid-HOG, Local Descriptor, Space-time Robustness.

Abstract: This article proposes a generalization of our approach REFA (Grand-brochier et al., 2011) to spatio-temporal domain. Our new method REFA3D, is based mainly on hes-STIP detector and E-HOG3D. SIFT3D and HOG/HOF are the two must used methods for space-time analysis and give good results. So their studies allow us to understand their construction and to extract some components to improve our approach. The mask of analysis used by REFA is modified and therefore relies on the use of ellipsoids. The validation tests are based on video clips from synthetic transformations as well as real sequences from a simulator or an onboard camera. Our system (detection, description and matching) must be as invariant as possible for the image transformation (rotations, scales, time-scaling). We also study the performance obtained for registration of subsequence, a process often used for the location, for example. All the parameters (analysis shape, thresholds) and changes to the space-time generalization will be detailed in this article.

1 INTRODUCTION

Today, digital imaging is becoming more prevalent in current applications of life. It is used for example to track, to localize, or to recognize. Scientists search and propose methods to acquire or create images, to edit content, or to extract all the information necessary for various applications. To give some examples, we can cite the 3D reconstruction, object tracking and the face recognition. These applications need data usually extracted with two tools: the detections of interest points and the local description. For 2D applications, we can cite methods such as SIFT (Scale Invariant Feature Transform) (Lowe, 1999; Lowe, 2004) and SURF (Speed Up Robust Features) (Bay et al., 2006), offering a complete system for the detection and local description of points. We proposed in 2011 the method REFA (Grand-brochier et al., 2011), to extract and characterize interest points with greater precision and a higher matching rate. The addition of temporal information is used to complete the analysis to study the mouvement of points in a video sequence. Processes such as localization or tracking required to use this type of data. Several methods offers this type of study, we can cite SIFT3D (Scovanner et al., 2007; Klaser et al., 2008) which is the generalization of SIFT, SURF generalized (Willems et al., 2008) or the coupling HOG/HOF (Laptev and Lindeberg, 2006; Laptev et al., 2007). To provide the best possible characteristic points of video for different space-time applications, we propose to generalize our approach REFA, making sure to remain as robust as possible against the various transformations existing between two video sequences (translations, rotations, scale changes, timescaling changes). We must also retain the various constraints that we set for our spatial method (robustness, matching rate and precision). All parameters of our new method REFA3D will be detailed in this article.

Section 2 presents briefly two space-time detectors and three characterizations of points, the method SIFT3D, SURF generalized and the couplig HOG/HOF. Additions, changes and parameters used for the construction of our new approach REFA3D are detailed in Section 3. To validate our method, we compare it with the SIFT3D and HOG/HOF, through various tests by implementing a number of transformations of data in section 4. We also propose results for the registration of subsequence.

2 RELATED WORK

Many approaches provide tools to extract and charac-

Grand-Brochier M., Tilmant C. and Dhome M..
REFA3D: ROBUST SPATIO-TEMPORAL ANALYSIS OF VIDEO SEQUENCES.
DOI: 10.5220/0003857203520357
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 352-357
ISBN: 978-989-8565-03-7

terize the interest points moving in time. For detection, we can cite Laptev and Lindberg (Laptev and Lindeberg, 2003), Dollar and al (Dollar et al., 2005) and Willems and al (Willems et al., 2008). The spatio-temporal description is generally based on the coupling, or the extension 2D+t, of existing methods such as SIFT (Lowe, 1999; Lowe, 2004) or SURF (Bay et al., 2006). A listing of these generalization was published by Wang and al (Wang et al., 2009). We limit our analysis to the SIFT3D (Scovanner et al., 2007; Klaser et al., 2008) and to the coupling HOG/HOF (Laptev and Lindeberg, 2006; Laptev et al., 2007)

Introduced by Laptev and Lindberg (Laptev and Lindeberg, 2003), Harris3D proposes a temporal generalization of the matrix of Harris, to obtain the tensor of structure:

$$\mathbf{M} = g_{\sigma,\tau} * \begin{bmatrix} I_x^2 & I_xI_y & I_xI_t \\ I_xI_y & I_y^2 & I_yI_t \\ I_xI_t & I_yI_t & I_t^2 \end{bmatrix} . \quad (1)$$

where $g_{\sigma,\tau}$ is the spacetime Gaussian function, defined by a spatial scale σ and by a temporal scale τ. Dollar and al. coupling in 2005 this approach with the impulsives responses of the temporal filters define by:

$$h_{ev}(t;\tau) = -cos(8\pi t)e^{-t^2/\tau^2} \quad \text{and} \quad h_{od}(t;\tau) = -sin(8\pi t)e^{-t^2/\tau^2}, \quad (2)$$

Willems and al. resume in 2008 the general idea of Laptev and Lindeberg to apply it to the hessian matrix and to create the hes-STIP (*hessian spatio-temporal interest point*) detector. Their goal is to propose a generalization of the SURF method, usually used for the images analysis.

To generalize the SIFT descriptor, Scovanner and al then Klaser and al, add it a 3D analysis model. Figure 1 illustrates their histograms HOG3D (Klaser and al.), by detailing the steps of construction.

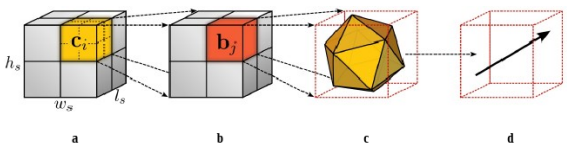


Figure 1: Various steps of the HOG3D construction: sampling of the mask of analysis (a et b), determination of the gradient orientation (d) in every sub-block with an icosahedron (c). (Klaser et al., 2008).

This approach consists to determine a region of 3D analysis, centred on the interest point. The mask is divided into $M \times M \times N$ blocks which is divided in turn into S^3 sub-blocks (Figure 1.a and 1.b). The orientation is determined with a regular polyhedron (Figure 1.c). Finally an histogram of oriented gradients is built on each $\mathbf{b}_j$.

A spatio-temporal extension of the SURF is proposed by Willems and al. The principle is to extend the Haar warvelet to a cuboid of size $s\sigma \times s\sigma \times s\tau$, where σ and τ are respectively the spatial scale and the temporal scale and s is a factor defined by the user. The descriptor is made up of the Harr wavelets responses x, y and t.

Laptev et al. (Laptev and Lindeberg, 2006; Laptev et al., 2007) combine different histograms to define the spatial and the temporal aspects. Their idea is to build a HOG with a spatial analysis 'classic' and pair it with a histogram of oriented optical flow (HOF) in order to have a temporal concept.

We presented various approaches of detection and local description, integrating a temporal analysis. The study of these methods allows us to extract the main advantages from it (stability, performances and invariances). We propose a generalization of our approach REFA (Grand-brochier et al., 2011) based on these diffrents tools and based on an ellipsoidal local exploration. So we detail in the next section, the modifications, the new parameters and the optimizations used.

3 METHOD

We propose a generalization of our method, to include space-time data to process video. To remain as invariant as possible to the various image transformations, our approach is divided into three parts: a hes-STIP detector (*hessian spatio-temporal interest point*), a local E-HOG3D (*ellipsoid histogram of oriented gradients 3D*) and an optimized matching. This section describes the different steps of our method and parameters used.

3.1 Detection

Proposed by Willems and al. (Willems et al., 2008), the hes-STIP is a generalization of the fast-hessian method (Bay et al., 2006), to include temporal data. This addition provides the following equation:

$$\mathbf{H}(\mathbf{x};\sigma,\tau) = \begin{bmatrix} L_{xx}(\mathbf{x};\sigma,\tau) & L_{xy}(\mathbf{x};\sigma,\tau) & L_{xt}(\mathbf{x};\sigma,\tau) \\ L_{xy}(\mathbf{x};\sigma,\tau) & L_{yy}(\mathbf{x};\sigma,\tau) & L_{yt}(\mathbf{x};\sigma,\tau) \\ L_{xt}(\mathbf{x};\sigma,\tau) & L_{yt}(\mathbf{x};\sigma,\tau) & L_{tt}(\mathbf{x};\sigma,\tau) \end{bmatrix} \quad (3)$$

Its construction is based on the interpretation of the hessienne matrix (equation 3) and particularly on two local scales: σ and τ. The first corresponds to the space exploration defined by the fast-hessien and the

second allows us to add a temporal analysis of the local information. To optimize this detector, we observe the influence of these two scales on the repeatability rate of our method. The results show that this rate is optimal for a spatial analysis following two octaves and a temporal exploration following four scales. The number of points is not the most significant for applications such as the homography estimation or objects recognition for exemple. On the contrary, good matchings precision increase strongly the quality and the performances, due to a lower number of outliers (false matchings). So we choose these criteria in spite of 7% loss of matched points.

3.2 Description

The local description of the method REFA is based on the use of histograms of oriented gradients following an elliptical mask. The addition of temporal data forces us to change our mask, transforming the ellipses in ellipsoids. In order to analyze the entire spatio-temporal information, we propose the mask shown in Figure 2, based on a sampling of the ellipsoidal neighborhood of the interest point. The latter is determined according to five levels of description (level -2 to level 2) combining 37 ellipsoids. For better visibility of the spatio-temporal aspect of our descriptor, we only display the centers of the ellipsoids in the illustration.

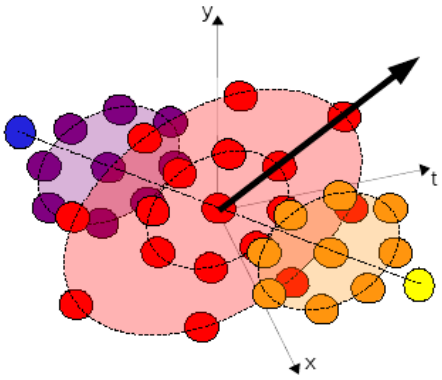


Figure 2: Representation of our analysis ellipsoidal mask, according to five levels of description.

The parameters of the ellipsoids are based on the scales (spatial and temporal) of local interest points. To increase the invariance to rotation, we adjust the mask analysis in two angles. The analysis of the matrix Harris3D (equation 1) introduced by Laptev and Lindeberg (Laptev and Lindeberg, 2003) to retrieve two angles θ and φ, shown in Figure 3.

The description of the method REFA is essentially based on the use of histograms of oriented gradients (eight classes). So the addition of temporal data forces us to change these histograms. Building on the work of Klaser and al. (Klaser et al., 2008), providing a generalization of HOG in space-time domain, we construct the following twenty classes. To do this, our histograms is based on an icosahedron (regular polyhedron) to optimize the distribution of such data. The choice of the class of the histogram based on the determination of the intersection of the gradient vector with one of the twenty faces of the icosahedron. To order our descriptor optimally, the face corresponding the first class of our histograms are readjusted according to the vector **v**. The latter corresponds to the combination readjustments shown in Figure 3.

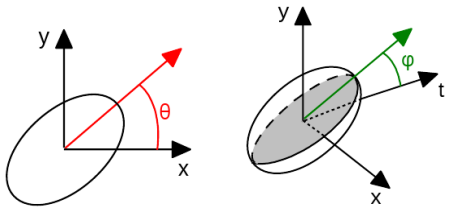


Figure 3: Illustration of spatial adjustement (left) and temporal (right) of an ellipsoid.

A final step is to saturate the values of the gradients, allowing us to increase the robustness to illumination changes. This process limits the influence of outliers characterized by high gradient values.

3.3 Matching

The goal is to find the best similarity (corresponding to the minimum distance) between descriptors des_{I_1} and des_{I_2} of two video sequences. Euclidean distance, denoted d_e, between two descriptors is defined by:

$$d_e(des_{I_1}(x_k,y_k,t_k), des_{I_2}(x_l,y_l,t_l)) = \sqrt{[des_{I_1}(x_k,y_k,t_k)]^T \cdot des_{I_2}(x_l,y_l,t_l)}, \quad (4)$$

where $(x_k,y_k,t_k) = \mathbf{x}_k$ and $(x_l,y_l,t_l) = \mathbf{x}_l$ represent the interest points respectively in the first and in the second sequence. The minimization of d_e, denoted d_{min}, provides a pair of points $\{(x_k,y_k,t_k);(x_{\tilde{l}},y_{\tilde{l}},t_{\tilde{l}})\}$:

$$\tilde{l} = \underset{l \in [\![0;L-1]\!]}{argmin} \left(d_e(des_{I_1}(x_k,y_k,t_k), des_{I_2}(x_l,y_l,t_l))\right) \quad (5)$$

and so

$$d_{min} = d_e(des_{I_1}(x_k,y_k,t_k), des_{I_2}(x_{\tilde{l}},y_{\tilde{l}},t_{\tilde{l}})). \quad (6)$$

To reduce the computation time, we generalize the decision tree used by the method REFA. The latter depends on the size of the data provided, its size is therefore $\mathbb{R}^{340}$ (seventeen histograms with twenty classes each). Regarding the selection threshold and method of removing duplicates, processes and parameters remain unchanged.

4 RESULTS

We are going to compare our method REFA3D with SIFT3D (Klaser et al., 2008) and the coupling HOG/HOF (Laptev and Lindeberg, 2006; Laptev et al., 2007). These two methods give good results for video analysis. We propose to study the matching rate and the precision of each of them. We will also study the subsequences registration.

4.1 Databases

The first database, noted *BSS*, is based on video extracted from an onboard camera. We then apply synthetic transformations such as translations (BSS_t), rotations (BSS_r), scale changes (BSS_{es}) or timescaling changes (BSS_{et}). Figure 4 illustrates these transformations.

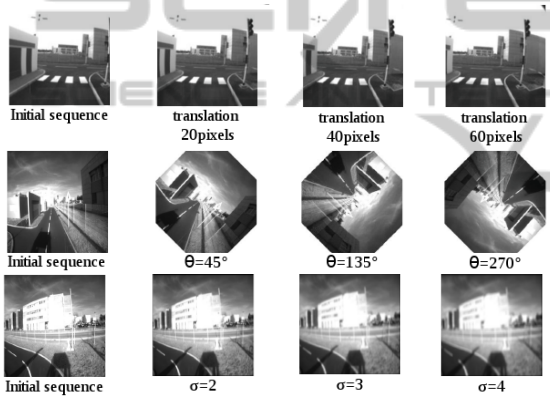


Figure 4: Examples of transformations (translations, rotationsby an angle θ and scale changes σ).

The second database, noted *BSR*, comes from the simulator ASROCAM (Malartre, 2011; Delmas, 2011), to create trajectories (*BRSs*, *BSRaq*, *BSRat*) in a virtual environment. Figure 5 shows an example of this database.

Figure 5: Example of an image sequence created by simulator ASROCAM.

4.2 Evaluation Tests and Results

4.2.1 Matching Rate and Precision

We propose to compare the matching rate as well as the precision of method REFA3D (blue), SIFT3D (yellow) and HOG/HOF (red). The matching rate is defined by the number of matches divided by the number of possible matches. The precision is defined by the number of correct matches divided by the number of matches performed. Figure 6 shows a synthesis of the results obtained.

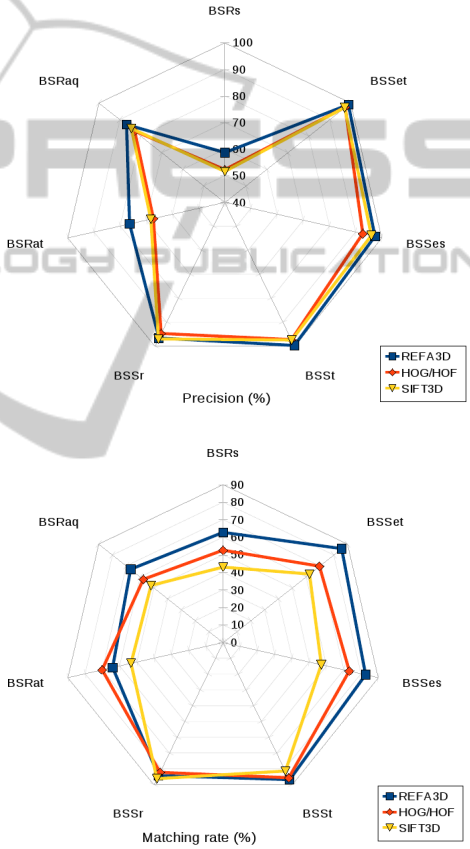
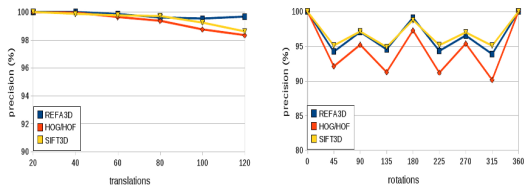


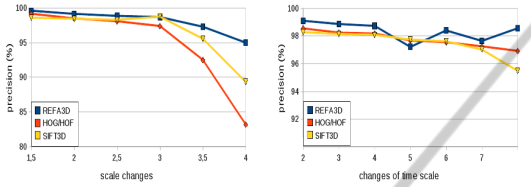
Figure 6: Summary of results for the spatio-temporal precision (left) and matching rate (right).

Given the different results, it appears that our approach has the best results in most cases. Its precision decreases for real changes, but remains higher than the HOG/HOF and SIFT3D. Our approach also provides a better overall matching rate, characterizing a description more relevant in the neighborhood. Finally our method REFA3D is more robust and stable for the various transformations considered. To detail the precision curves of different methods, we propose Figures 7 and 8.



(a) (b)

Figure 7: (a) Precision rate for translation (in pixels) and (b) precision rate for rotations (in degrees).



(a) (b)

Figure 8: (a) Precision rate for scale changes and (b) precision rate for timescaling changes.

Concerning transformations studied, our method has generally a higher precision than other methods or similar to SIFT3D in the case of rotations. The stability also enables us to conclude that a better robustness of our approach. Nevertheless, these results are based on various tools (optimization, threshold) involving a slight decrease in the number of matched points.

4.2.2 Subsequences Registration

We propose a study of the subsequence registration. First we analyze three trajectories: a straight line, a curve and a subsequence simulation. Table 1 show the precision "*P*", the number of matches "*Nm*" and the frame rates are registrated "*Fr*", for three methods compared. It appears that our approach has a registration generally with a better precision of matches and the rate of registered images is greater. Our approach therefore presents a more relevant description of the scene. The only disadvantage is the decrease in the number of matching.

We propose a final test by implementing readjustments of five subsequences in an obstacle avoidance. Figure 9 illustrates the five stages of obstacle avoidance, and subsequences associated. Table 2 shows the results ("*P*" for the precision in percent and "*Fr*" for the frame rates are registrated in percent) of REFA3D methods, SIFT3D and HOG/HOF. The analysis of these results shows that our approach gives a precision rate and registered images generally higher than those of the methods compared. Only SIFT3D presents, for the subsequence ss5, a higher precision. Our method also has better stability, represented by decreases lowest observation criteria. With the performances obtained by our method, it would be interesting to consider the use of these data in a different process of realignment of the vehicle on its nominal trajectory. The matches extracted by our approach would estimate, frame by frame, the homography and thus to calculate the various parameters of registration to provide the localization system.

Table 1: Results for the registration of subsequences for our method REFA3D, the coupling HOG/HOF and the method SIFT3D.

	P	*Nm*	*Fr*
	REFA3D		
Straight line	99.8%	204	100%
Curve	97.6%	155	97.6%
Simulator	97.4%	237	98.3%
	HOG/HOF		
Straight line	99.2%	212	99.6%
Curve	96.9%	178	95.3%
Simulator	94.8%	256	92.5%
	SIFT3D		
Straight line	98.7%	284	98.1%
Curve	97.2%	247	97.8%
Simulator	95.4%	294	93.2%

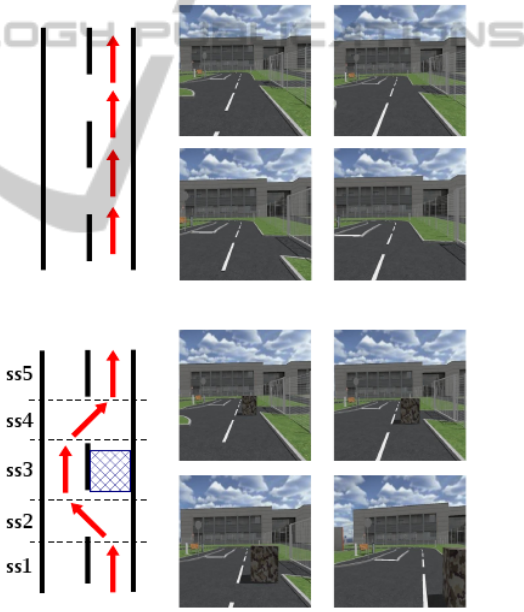


Figure 9: Samples of the initial sequence and those with an obstacle avoidance (split into five subsequence).

Table 2: Results for the registration of subsequences in an obstacle avoidance.

	REFA3D		HOG/HOF		SIFT3D	
	P	*Fr*	*P*	*Fr*	*P*	*Fr*
ss1	99.4	100	98.7	99.5	99.2	100
ss2	91.3	95.6	89.2	90.3	86.7	93.3
ss3	79.1	87.6	67.3	72.3	71.2	81.3
ss4	85.4	92.2	79.6	84.4	81.3	88.4
ss5	94.7	97.3	91.3	91.5	95.1	95.6

5 CONCLUSIONS

We propose in this article a space-time generalization of our method REFA. To do this, we use the detector hes-STIP, which has the highest repeatability rate for this type of analysis. The optimization that we bring on the limitation of exploration scales (spatial and temporal). The mask of analysis is also modified to add the time component in the histograms. The ellipses are converted into ellipsoids and we use five levels of description (Figure 2). Adding a temporal adjustment results a stable three-dimensional exploration of the sequence. To validate this space-time generalization, we first propose several tests based on sequences from a real camera and a simulator. The results show that our approach generally gets the best precision. We also observe a better stability and a higher matching rate. In a second step, we study the registration of subsequences. This type of process is used to provide space-time informations of the object (localization, trajectory for example). Our method performs best for the precision and the rate of registered images.

Our future prospects is the integration of our approach REFA3D in intelligent vehicles. Our goal is to improve again and again the precision of our method, for the vehicules to be more reliable and secure. An other prospect is to export our descriptor to three-dimensional field to use it in medical imaging.

REFERENCES

Bay, H., Tuylelaars, T., and Gool, L. V. (2006). Surf : Speeded up robust features. *European Conference on Computer Vision*, pages 404–417.

Delmas, P. (2011). *Gnration active des dplacements d'un vhicule agricole dans son environnement*. PhD thesis, University Blaise Pascal - Clermont II.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. *IEEE International Conference on Computer Vision*.

Grand-brochier, M., Tilmant, C., and Dhome, M. (2011). Method of extracting interest points based on multi-scale detector and local e-hog descriptor. *International Conference on Computer Vision Theory and Applications*.

Klaser, A., Marszalek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. *British Machine Vision Conference*, pages 995–1004.

Laptev, I., Caputo, B., Schuldt, C., and Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, 108(3):207–229.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. *IEEE International Conference on Computer Vision*, 1:432–439.

Laptev, I. and Lindeberg, T. (2006). Local descriptors for spatio-temporal recognition. *Computer and Information Science*, 3667:91–103.

Lowe, D. (1999). Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision*, pages 1150–1157.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Malartre, F. (2011). *Perception intelligente pour la navigation rapide de robots mobiles en environnement naturel*. PhD thesis, University Blaise Pascal - Clermont II.

Scovanner, P., Ali, S., and Shah, M. (2007). A 3-dimensional sift descriptor and its application to action recognition. *ACM Multimedia*.

Wang, H., Ullah, M., Klaser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference*.

Willems, G., Tuytelaars, T., and Gool, L. V. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector. *European Conference on Computer Vision*, 5303(2):650–663.