Optical Flow estimation and Global motion estimation in the image plane with RANSAC algorithm

Meaza Eyakem Gebreamlak, Sebastián Cajas Ordoñez,
*University of Bordeaux.

I Introduction

Nowadays, the number of motion applications have severely increased, ultimately improving the way we give to machines the ability to detect objects and understanding movement. Gestalt had rooted the origins of perceptual grouping, with the goal of understanding how humans detect objects, creating a description for moving objects and that today have a great number of applications on computer vision and deep learning. [1]

During the present work, several evaluation metrics of information content of image motion sequence are studied, including Mean Squared Error (MSE), Peak Signal-to-Noise Radio (PSNR), entropy and optical flow estimation methods using openCV modules over two different videos with a $deltaT$ of 10 and 100. The second part is focused on implementing the optical flow, the compensated frame and different curves demonstrating the image error differences between the two $deltaT$ values. Ultimately, the Global motion estimation (GME) is implemented using those retrieved optical flows. The residual motion is also depicted and analyzed. The code is tested over all the videos and on $CREMI$ servers.

II Literature

Amongst the multiple applications of motion estimation of objects, it is particularly useful on object detection. Starting from background subtraction, which consists in extracting moving objects by subtracting a static background, to shot boundary detection methods or motion segmentation techniques. The last one, created with the goal of extracting objects using object segmentation. Other techniques involved feature based methods, which depend fundamentally of visual features, and secondly, the direct methods, which relate deeply with the inner properties of the images [2] [3]

Motion compensation error: This is calculated with the function $computeErrorImage$, which represents the motion compensation error for mono-resolution frames difference. This should generate an image in three colors: gray (128), black (0) and white (255); gray, when there are no changes between pixel positions; black or white colours when there is a transition between white to gray, white to black, gray to white or gray to black.

Motion Compensation error will always be present because of three different factors

- *Problem of existing occlusion*: The difference between consecutive frames with moving object will create undiscovered pixel values, also known as parasite regions, which will be seen as unmet sections from which there is no information about.
- *The aperture problem* or the optical flow constrained has 2 variables, but at the same time it's a differential equation of second order, therefore there are infinite possible solutions.
- *Sensitivity to noise*, due to the usage of derivatives, there is a lot of noise.

A. Evaluation metrics

Mean Squared Error - MSE: MSE calculates the squared difference between the current and previous image.

$$MSE = \frac{1}{N * M} * \sum I(p,t) - I(p, y - delta)^2 \quad (1)$$

Peak Signal-to-Noise Ratio (PSNR): Based on the given MSE, computes the Peak Signal to Noise ratio based on a real factor, max_pixel and is inversely proportional to MSE.

$$PSNR = 20 * \log \frac{Max\_value}{MSE} \quad (2)$$

Therefore, the plot should look like an inverted and re-scaled version of MSE. To avoid zero denominator, MSE is set to 100 when MSE is near small values, therefore reaching the value. To avoid zero-division on this equation, minimum value to obtain would be therefore

$$20 * Log10(255/100) = 8 \quad (3)$$

Entropy: The goal of using entropy at this stage is to analyze the information quantity changes between consecutive frames, in the following way:

$$Ent(X) = \sum p(x_i) \log_2 * p(x_i) \quad (4)$$

If we consider entropy zero, this will mean that there is a lot of information between the considered consecutive images. If we consider as well a high value of MSE, that would mean that there is motion and therefore a lot of information, which will yield to lower entropy, as the probability distribution will increase. If we consider the opposite case, when entropy is 1, this would mean that the probability distribution for the current image is very low, in other words, will have a low amount of information and therefore high entropy.

As a conclusion, we will use entropy of an image because it tell us how much movement information we have available. In case of images, intensity depth is 8 bits, meaning that for digital pixels we will have 256 possible values, therefore the maximum value we can obtain for entropy is 8.

Image Error It will be computed from the function $ErrorImage$, which will depict motion in three escenarios: gray color when no motion is present and black/white color, when motion is present. This will be done by subtracting 2 images or flows (estimation) and add 128 pixel value which represents gray color scale when the difference between 2 pixels equals zero. Then this value is set to have a maximum value of the subtraction plus 128. From this we can make the following conclusions, which will be handy during the analysis of the images:

- If we have that the pixel-wise subtraction between the consecutive images is equal to 0, then adding 128, which will yield to gray color. This means that if there was no motion this will be depicted with gray color.
- If the pixel-wise subtraction yields 255, that is, when $image1 > image2$, then this will mean that the pixels in image 1 was white and the motion moved them to white.
- if the pixel-wise subtraction yields -128, then this would mean that there was a value change from 128 to 255 values, that is, from no moving representation to one that has movement. Adding 128 value to this result will yield to a zero pixel value, meaning that this movement will be depicted as black color.

Optical Flow computation

Flow is based on the fundamental hypothesis of motion estimation, which states that the intensity along a trajectory sums up to zero. It will create a 2nd order differential equation for which we intend to minimize and optimize to find the velocity vectors $W(u, v)$ with the best possible values, based on an iterative methods. However it will contain certain natural error, "ill-posed error" based on occlusions, sensitivity to noise and problem of apperture. During the solution of the Aperture problem constrained:

- The dot product between the gradient of the image and the velocity vector $W(u, v)$ will be equal to $-It$, therefore since this is a dot product, we will have a relationship between the image gradient and velocities, which are both considered as velocity vectors, with correspondent x and y components (norm and tangent), so the $Wnorm$ will be parallel to y-axis if the image gradient in terms of its tangent and norm, because the gradient of the y component is parallel to the norm component if Image I, then the whole expression is self-contained on the tangent vector, as the norm is canceled out as it is orthogonal to the local contour. At the end what will remain is the normal optical flow.
- W can be determined with several methods, one is the "Estimation of Optical Flow with regularization Horn and Shunk method". This method will contain the regularization term times a constant plus the first term, which is the aperture constrained equation. The value of the regularization constant alpha will determine the smoothness of the flow. If Alpha is big, then there will be a high smoothing, but if alpha is 0, there there will not be smoothing at all.

CompensatedFrame: It will create the Estimated Flow which will be used to compute the MSE, taking the linear interpolation between the previous image and the flow.

III Results and Analysis

A. Analysis on the first video

The first video sequence that is used to analysis the result is the vtest.avi video see fig 1. Two value which are 10 and 100 are given to the delta parameter for the analysis purpose. DeltaT is a parameter which indicates the difference of time between current and previous frame.

The video shows many people moving and crossing the street. Some of the main observations we can look from this video is there are many occlusions during the duration. Some of them are the existence of the standing light pole that occludes the people who passes behind it. Second one is when the people cross pass each other and the one person occludes the other. Also, when people are greeting the other person tends to occlude his fellow. When people are walking side by side one of the people half side gets occluded all the time.

The second obstacle in the image is there are multiple shadows of the moving people in the image. Which would affect the quality of the motion estimation in great degree.

Fig 1 shows the **Current Frame** on the left side the **Compensated frame** obtained using delta value of to 10 on the center and Compensated frame obtained using the deltaT value of 100 is on the right on *vtest* video.

As we can depict from figure 1 compensated frame(the frame

Fig. 1. **Current Frame** (left), **Compensated frame with delta equal to 10** (center) and **Compensated frame with deltaT equal to 100** (right) on *vtest* video.

which is compensated using the optical flows) is blurry with a lot of artifacts and noises. There are many objects that are either deformed or lost totally. In addition to that, many pixels that were supposed to be still are moving like in the case of the shadows they are almost always moving to certain degree with the person.

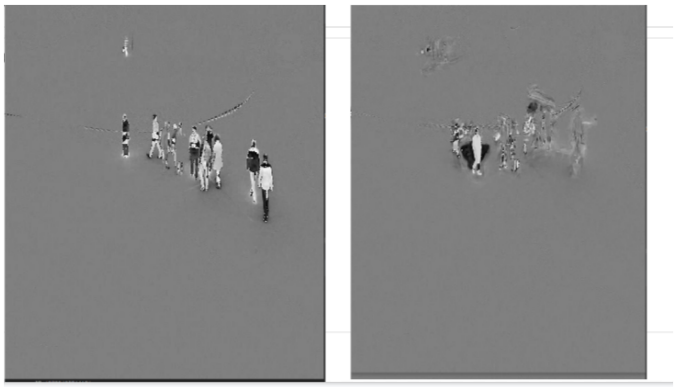
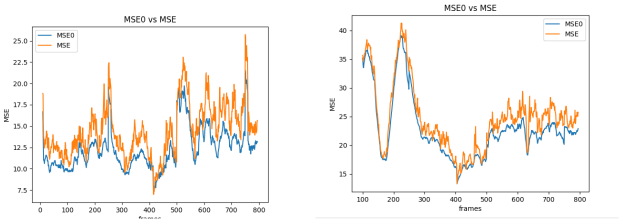


Fig. 2. **Imerr0** (left) and **Imerr1** (right) for vtest video

Figure 2 shows the images errors comparing with the previous frame and the compensated frame consecutively.

The image error is calculated by comparing the intended frames. In $ImErr0$, it is between the previous and current frame and in $ImErr$ it is the between the compensated frame and current frame. Gray means it is static, black indicates it is moving in the negative direction and white indicates it is moving in the positive direction. As we can observe, the image error for the compensated image is not good with a lot of artifacts and parasite images. On the case of *vtest* sequence, it is due to occlusions and shadows.

This result is due to the sensitivity problem of motion estimation. As it is depicted in the previous section the motion estimation problem do not have unique solution. During compensating the frame, when it encounters weak noise, it provides large difference in the prediction. In addition to this, since there are numerous objects which



(a) delta 10 (b) delta 100

Fig. 3. MSE calcuated using different on *vtest.avi* video

are occluded many times, it is reasonable for the output to turn unsatisfactory. The image error for the current frame and previous frame provides good results with good edge detection and less artifacts.

MSE metric is used to evaluate the information of the motion sequence.The values obtained using delta value of 10 and 100 are shown in figure 3. The highest peaks of the plot indicates that there is big difference between the compared frames. This situation arises due to the fast movement of objects and in greatest extent during appearance and disappearance of an object. As a result of this the curve is not constant during the whole duration as the frames movement varies greatly in the video.

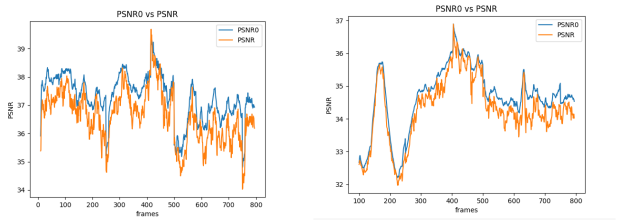
As it can be seen from figure 3, the *MSEs* calculated between the previous frame and current frame (which is denoted by *MSE0*) are less than the *MSEs* calculated between the compensated frame and current frame. The larger values indicates that there is more information content of motion sequence than the smaller values.

Different values of deltaT are used to evaluate the effect of the parameter. At the beginning of the video there is high time difference between the consecutive frames which yields high *MSEs* value for the 100 value of deltaT.

PSNR is the inverse of MSE as it is calculated by dividing the noise value. Therefore, means the higher PSNR describes low information content of the motion sequence. Therefore, from fig 4 One can observe the PSNR0 (PSNR calculated between previous and current frame) is greater than PSNR (psnr calculated b/n compensated and current frame).

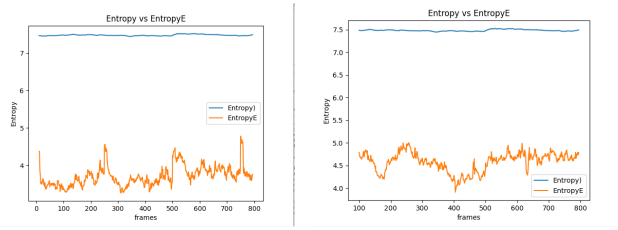
The PSNR value obtained using deltaT 100 yields low value at the beginning of the picture. This result is expected since there is high time difference between the consecutive frames which yields high error or noise value. Therefore, if there is large noise in the flow the corresponding PSNR value will be lower.

Entropy is used to measure the information content of the flow. The entropy of the gray frame and previous frame displayed in figure 5 has a high value while the entropy value for the flow obtained with the compensated and current frame



(a) delta 10 (b) delta 100

Fig. 4. PSNR calcuated using different on *vtest* video



(a) delta 10 (b) delta 100

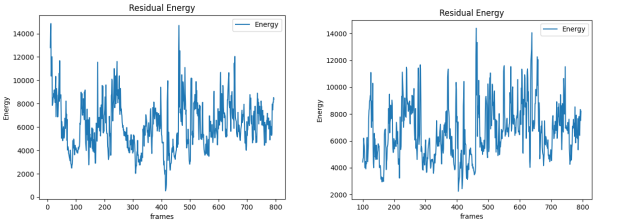
Fig. 5. ENTROPY calcuated using different on *vtest.avi* video

has relatively low. The gray frame also has a constant value while the entropy for the compensated frame has a cure with a lot of swings.

The **residual energy** between the original optical flow vector and the global estimation is presented in figure 6. The curves of the plot have many swings indicating that there is a big difference in the motion of consecutive frames in the video sequence. High values indicate that there is big error in the consecutive errors and low value describe there is a less error between consecutive frames.

B. Analysis on the second video

The second video sequence entitled *People convergence sequence*, considers a single person moving across the street. On it, different illumination factors are present, in particular the sky color and and the water reflection on the river. There is also a high contrast between the different objects due to highly distinguishable pixel values identifying different regions on the same background, meaning that these objects are prone to contain noise due to extra noise detection.



(a) delta 10 (b) delta 100

Fig. 6. Residual Energy calcuated using different on *vtest.avi* video

On figure 7, the current frame and the compensated image is compared when *deltaT* is 10 and 100. On it, it is possible to observe optical artifacts present on the image. This deformation is mainly created due to the ill-posed problems stated earlier during the flow extraction, where multiple previous background pixel values were unknown and therefore when interpolating the previous frame and the flow with linear interpolation, an optical deformation or motion artifact will be shown, in particular on the areas where the person's limbs are in constant movement.

Fig. 7. **Current Frame** (left), **Compensated frame with delta equal to 10** (center) and **Compensated frame with deltaT equal to 100** (right) on *Person Convergence* video.

The figure 8 shows *ImErr0* and *ImErr* for delta 10 and 100. As commented on the literature section of this report, given the fact that there is a high content of edges detection due to the main features of the video sequence, *ImErr0* in general is highly more sensitive than *ImErr*, that is, it should contain less amount of error due to the ill-posed problems. *ImErr* Should not be equal to *imErr0*, because on this case we are comparing the current image (gray) with the compensated flow, which by definition will always have motion artifacts and sensitivity to noise, therefore the two graphs will always be different.

On figure 9, a comparison between the flow, Global motion estimation (GME) is calculated when $deltaT$ is 10 and 100. The flow shows the velocity vector $W(u)$ graphically. From it, we can observe the velocity vectors are constantly changing its magnitude, in particular around the moving object, demonstrating that the vectors directions are based on how the motion happens. This computation is performed by the function $draw\_flow$, which plots the current image and at the same time, will plot the calculated flow.

Secondly, the GME is be calculated over the entirety of the image, using the original gray image and the one adding the flow to it. Afterwards, the homography for the estimated and consecutive frame is calculated, and finally, a perspective transformation is applied to the source image using the homography. The final global motion is calculated as the subtraction of the source and the given transformation.

As a result, it can be seen the GME represents the entire movement of the image, as the vectors values remain with constant change all at the same pace; as been compared with the flow graph, the main difference can be seen as a a change

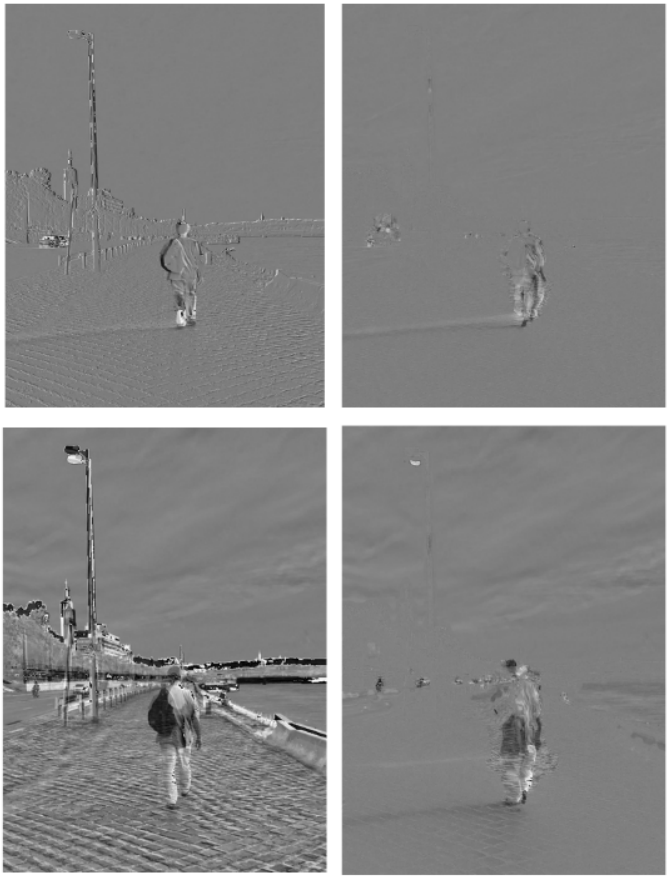


Fig. 8. **ImErr0 - deltaT of 10** (upper-left), **ImErr - deltaT of 10** (upper-right), **ImErr0 - deltaT of 100**(lower-left),**ImErr - deltaT of 100** (lower-right), for *Person Convergence* sequence.

on vectors directions is centralized around the moving object region for the flow graph, while for the *GME* remains constant.

When the error between the flow and *GME* is calculated, we can observe that the main changes occur around the moving object. There are other present factors on the graph as well, mainly due to changes in illumination. But most importantly, when *deltaT* is increased to 100, we can observe the high increase of the velocity vectors in the flow graph for the persons convergence sequence, which is highly present in particular where the motion happens. This means that using this frame difference, more background details can be neglected as they are too small, but also it would be necessary to find an optimal *deltaT* factor that minimizes as best as possible this noise without compromising the moving object quality, which in this case is affected.

The mean squared Error (MSE) for *deltaT* equal to 10 and 100 is depicted on figure 9. On this case, we can observe that *MSE0* is always bigger than *MSE*, demonstrating that the amount of edge detection between the previous and current frame is bigger than for the compensated frame and gray. Again, the big difference between 2 graphs is that *MSE* has less error even for the positions with strongly differences in pixel movement, demonstrating that it is mostly smoothing the previous image at each frame and even decreasing its values over frame. This explains why on the compensated image

Fig. 9. **Flow** (upper-left), **Global Motion Estimation**(upper-center), **Global Motion Estimation Error**(upper right) for *People convergence* sequence using $deltaT$ of 10 on the fist row, and using $deltaT$ for the second row.

for *imErr*, there is less border detection and mainly focuses on the person. However this detection is vain, as there are multiple artifacts around the person as well. Regarding the difference for $deltaT$ of 10 and 100, we can see that when its 10, there is a higher edge detection, which is traduced as higher MSE values, which vary at a broader range. Whilst if $deltaT$ increases to 100, edge detection is highly reduced, which explains why the standard deviation is reduced and *MSE0* starts to reassemble to *MSE* standard deviation, but with a different mean.

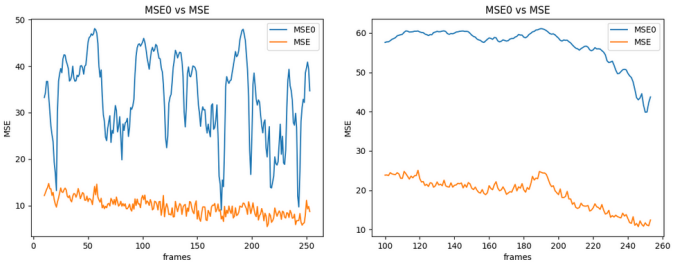


Fig. 10. **MSE for delta 10** (left), **MSE for delta 100** (right) for *People convergence* sequence.

The figure 11 represents the graphs for $PSNR$ in $deltaT$ 10 and 100. As it is the inverse relationship with MSE times a variable factor, we should obtained a reversed version with different amplitude of graphs for MSE. In the case of $deltaT$ of 10, we observe the same behaviour as in $MSE0$ but inversed, high standard deviation compared to $PSNR$ due to the same edge artifact detection. On $deltaT$ of 100, the standard deviation is highly reduced and is even smaller than in 10, because we have an scaling factor on $PSNR$ equation which reduces the standard deviation, but mantains the mean.

The entropy depicted in figure 12 will for either $deltaT$

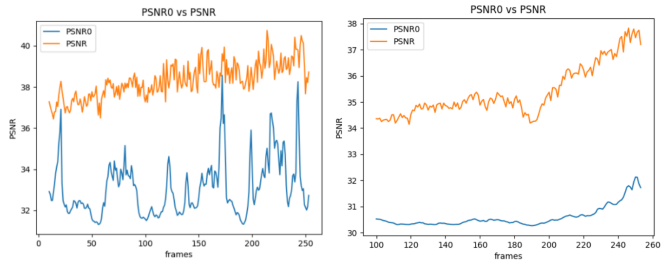


Fig. 11. **PSNR for delta 10** (left), **PSNR for delta 100** (right) for *People convergence* sequence.

equal 10 or 100 remains constant at its maximum level of 8, which corresponds to the maximum pixel value that can be stored for images. This is because most of the image at the beginning is static, since there is no movement, the probability distribution will remain very low or near zero, yielding therefore to high entropy. When movement is detected, the probability distribution for each pixel will severely increase, yielding to an decrease on probability. We can see that between $deltaT$ 10 and 100, the area of the second is higher across number of frames, this means that for a higher frame difference, there is higher motion detection.

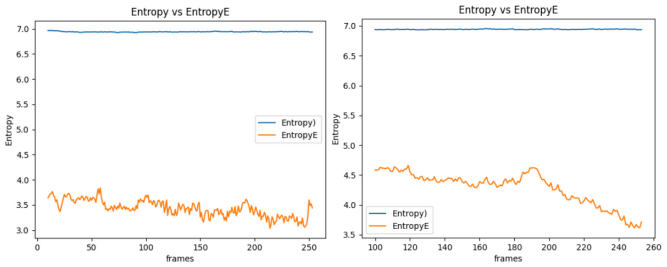


Fig. 12. **Entropy for delta 10** (left), **Energy for delta 100**(right) for *People convergence* sequence.

The energy depicted on 13 shows a lower standard deviation when $deltaT$ equals 10 than when its 100, since energy is equivalent to de subtraction of GME and flow, we can deduce from this that the higher the time difference between frames, the higher the energy values, because the error or the difference between flow and GME will be much higher and therefore, this results on a highly increase of the energy. On this sequence, for $deltaT$ equal 100, increases up to 3 times more than $deltaT$ of 10

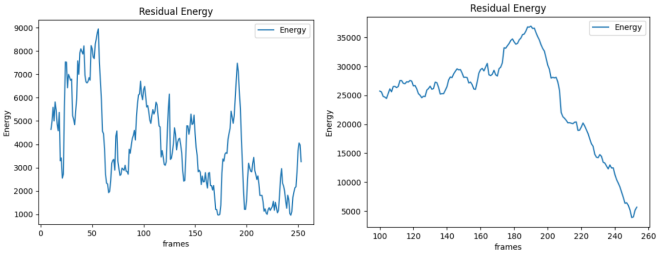


Fig. 13. **Energy for delta 10** (left), **Energy for delta 100** (right) for *People convergence* sequence.

IV Conclusions

- In general, we were able to observe the many-possible implications of the sequence, involving the context where it was recorded, the camera position and environmental positions, which will have an enormous effect in the calculation of the optical flow, the compensated frame (estimated frame for the current frame) and therefore the statistical metrics which were required to test ($MSE, PSNR, Entropy, Energy$), which will vary accordingly relative to these conditions, but will always keep a close relationship between each other: The calculation of the optical flow is highly entangled with the 3 ill-posed problems, meaning that there will always be an error to compensate, which is inherently related to the sensitivity to noise of the present video, its challenges (on this case, mainly occlusions and changes on illumination), aperture problem for multiple possible solutions and sensitivity to small values due to usage of derivatives during optical flow calculations.
- If we compute $ImErr - ImErr0$, we can obtain the loss of information between the gray and the flow, demonstrating that the flow contains more motion artifacts. Finally, the main difference between $deltaT$ of 10 and 100, is the amount of local edges, which is highly visual on the captured frame. The bigger the $deltaT$ values, the bigger the deformation will be present on the image, this can be explained because the bigger the time difference between the previous and current frame, the higher the pixel content is lost and therefore this can be seen as with a higher optical effect.

References

[1] Hörhan, M., Eidenberger, H. Gestalt descriptions for deep image understanding. Pattern Anal Applic 24, 89–107 (2021). https://doi.org/10.1007/s10044-020-00904-6

[2] Philip H.S. Torr and Andrew Zisserman: Feature Based Methods for Structure and Motion Estimation, ICCV Workshop on Vision Algorithms, pages 278-294, 1999

[3] Rui Xu, David Taubman and Aous Thabit Naman, 'Motion Estimation Based on Mutual Information and Adaptive Multi-scale Thresholding', in Image Processing, IEEE Transactions on , vol.25, no.3, pp.1095-1108, March 2016.