Video Completion as Noise Removal

Visesh Chari, C. V. Jawahar and P. J. Narayanan
Center for Visual Information Technology
International Institute of Information Technology
Gachibowli, Hyderabad - 500032
Email: ukvisesh@students.iiit.ac.in, {pjn,jawahar}@iiit.ac.in

***Abstract*—Video completion algorithms have concentrated on obtaining visually consistent solutions to fill-in the missing portions, without any emphasis on the physical correctness of the video. Resulting solutions thus use texture or image structure based cues and are limited in the situations they can handle. In this paper we take a model based signal processing approach to video completion [1]. Completion of the video is then defined as satisfying the given model by detecting and removing the error (selected parts of the video to be replaced). Given a probabilistic model, video completion then becomes an unsupervised learning algorithm with the input video giving a "noisy" version. Dense completion is the automatic inferencing of the "noise-less" or "true" video from the input. This approach finds a solution that satisfies visual coherence and is applicable to a wide variety of scenarios. We demonstrate the efficacy of our approach and its wide applicability using two scenarios.**

I. Introduction

The problem of video completion or inpainting deals with correction or replacement of selected parts of a video from content taken from the rest of the video. The recent need to automatize the restoration of various degraded videos has generated a lot of interest in this field [2]. The two fundamental sub-parts of the problem involve (i) identifying the parts of the video to be replaced and (ii) identifying the appproprate parts of the video to replace with. Solutions to both these problems involve registration of the frames of the video with respect to each other. Setting this problem in a signal processing framework is straightforward. The parts of the video that need to be replaced may be termed the noise in the signal, which is in turn represented by the content of the video. This allows us to borrow from the rich literature present in model based signal processing [1]. Different models (registration between frames) may be used to describe the original signal, which give rise to different solutions.

Approaches to video completion may be categorized into two frameworks.

a) Non-parametric approaches: When a video is modeled as a space time collection of texture and edge information, non-parametric methods like texture sampling [3] may be applied to remove noise in the video, assuming noise is already isolated or detected [4], [5]. Probabilistic approaches in this direction involve [6], where large number of images are used to learn texture patches called "epitomes" that are later used in the synthesis of videos. Other solutions based on optical flow [7] or partial differential equations describing the edge structure [8] have also been proposed. Interesting approaches involve [9], where the authors use cyclic motions in the foreground or background to correct the noise present in the videos. Such an approach is useful either when there is a large amount of data (texture or edge patches) available, or when the amount of noise is small compared to the overall signal (high signal to noise ratio (SNR)). This corresponds to scenes that are primarily affine or when the motion in the scene is orthogonal to the camera view.

b) Physical model-based approaches: The other approach involves modeling the physical phenomena occurring in the video. Thus, it involves modeling a video as the time-sequenced projections of a 3D event. Such an approach, however, has the advantage of being able to not only *correct* the noise in a video, but also *detect* it, making the whole completion process autonomous. [10] describe such a framework where the static background may be extracted from the scene in an affine setting. They assume the dominant optical flow in the scene to belong to the background, and extract foreground by clustering out flows that do not match the dominant one. The extracted optical flow of the background is then used to fill the removed pixels. In [11], a PDE based approach is used to detect and remove specularity from images and videos. The formulation of the PDE depends on the type of image source, texture information present in the scene etc. However, only specularities are handled in this formulation.

Thus approaches mentioned uptil now lack in two respects. Either they concentrate on producing visually appealing results, neglecting the "correctness" of the results obtained, which severely restricts their applicability. For example, a video with multiple moving objects occluding each other is challenging because of the huge number of parameters that need to be considered. Or the scenarios in which these algorithms work are restricted (affine, specularity). Ideally, however, we would like our formulation to be independent of such constraints. Thus it is desirable to have an algorithm that would a) automatically identify what parts of a video need to be replaced based on some cost function in a general 3D scene, b) be able to find physically correct patches to replace missing parts of a video, and c) inpaint objects of various and varying sizes and shapes.

To the best of our knowledge, the first two characteristics mentioned above are not handled by current video completion algorithms except [10], [11], and the third characteristic poses problems to texture based approaches since non-parametric sampling techniques are sensitive to scale changes. Thus, scenarios are typically restricted to scenes where objects to be removed do not change scale significantly, or the camera movement is restricted.

We present a novel approach to the problem, by defining inpainting without manual interventaion as an unsupervised learning problem. A single cost function specifies what parts of the video need to be replaced, and also identifies the appropriate physically correct replacement patches. In this paper, we consider two types of cost functions based on registration between views, and show how one helps in removing dynamic objects and the other in removing non-planar objects from a given scene.

This approach is along the lines of some recent papers [12], [13] which take a learning approach towards geometric problems. In dynamic mosaicing [12], the basic problem is to generate a mosaic of the static scene in the presence of dynamic objects, which are treated as noise and need to be removed. This requires registration across scenes, as in our case. Super-resolution [13], on the other hand tries to estimate a high resolution video from a low-resolution one, and the problem is posed in a supervised learning framework.

In subsequent sections, we first present our approach to the problem. We show how video completion can be posed as an unsupervised learning problem in the presence of noisy data. The definition of noise determines the exact structure of the algorithm, and two scenarios are presented to illustrate our approach. The first one consists of a planar object affected by illumination artifacts, and the second scenario consists of people moving in a 3D environment. Both situations are extremely challenging from the current stand point of the video completion community. Impressive results in both scenarios illustrate the efficacy of our approach.

II. Inpainting as outlier replacement

A video is a sequence of images produced by the projection of a 3D world onto a camera. Our interest lies in identifying and removing certain parts of this world from the images making up the input video. Thus, the input video and the true video may be represented as follows.

$$\mathcal{V}_I = \{\mathcal{I}_{I1}, \ldots, \mathcal{I}_{IN}\} \tag{1}$$
$$\mathcal{V}_T = \{\mathcal{I}_{T1}, \ldots, \mathcal{I}_{TN}\} \tag{2}$$
$$\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_N\} \tag{3}$$
$$\mathcal{V}_T = \mathcal{P}(\Psi) \tag{4}$$
$$\mathcal{V}_I = \mathcal{P}(\Psi) + \eta \tag{5}$$

where $\mathcal{V}_T$ represents the video to be inferred, and $\mathcal{V}_I$ represents the input to our algorithm. $\mathcal{P}$ represent the camera viewpoints, Ψ is a function of the 3D world, and η represents noise, which may be blacked out or degraded pixels, illumination artifacts, effects of jerky camera motion or occlusion, occluding or dynamic 3D objects etc. Notice how our approach does not need to know the complexity of the occlusion. The only assumption we make is that the whole of Ψ is visible without noise, in parts, somewhere in $\mathcal{V}_I$. Also, the definition of Ψ changes with the problem under consideration. Finally, the solution boils down to the estimation and replacement of this noise, with data from the video. Thus, we pose the problem of video completion as extrapolation of a function after fitting the same to input data in the presence of noise. Several approaches to this problem may be found in the machine learning literature [14], and here we take the approach of maximum likelihood estimation.

The solution proceeds in three steps. Since we follow a maximum likelihood formulation, we first need to evaluate the probability that a candidate hypothesis $(\hat{\Psi}, \hat{\mathcal{P}})$ produces the given video $\mathcal{V}_\mathcal{I}$. Maximizing this likelihood over the space of $(\Psi, \mathcal{P})$, gives us the best hypothesis. Noise is then represented as outliers of the model $(\Psi, \mathcal{P})$, and is removed by extrapolating the model at appropriate points.

c) **Estimating $\Psi, \mathcal{P}$::** Given a candidate hypothesis $(\hat{\Psi}, \hat{\mathcal{P}})$, the total probability of the observed video V_I is given as

$$\begin{aligned} p(V_I|\hat{\Psi}, \hat{\mathcal{P}}) &= \prod_i p(I_{Ii}|\hat{\Psi}, \hat{\mathcal{P}}) && (6) \\ &= \prod_i p(I_{Ii}|\mathcal{P}_i(\hat{\Psi})) && (7) \end{aligned}$$

where $\mathcal{P}_i$ represents the camera matrix corresponding to the i^{th} view. Thus, the probability that $\hat{I}_{Ti} = \mathcal{P}_i(\hat{\Psi})$ represents the "true" i^{th} view can be determined by maximizing the above equation. We assume $P(I_{Ii}|\hat{I}_{Ti})$ to be a Gaussian over pixel differences, with each pixel's contribution being independent of the others.

$$p(I_{Ii}|I_{Ti}) = \prod_{\forall x,y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{I_{Ii}(x,y) - \hat{I}_{Ti}(x,y)}{2\sigma^2}\right) \tag{8}$$

In other words, every individual frame of the best hypothesis should explain the current image to the best possible extent. We proceed to minimize the corresponding log-likelihood function

$$\begin{aligned} \mathcal{L}(I_{Ii}) &= -\sum_{\forall x,y} (I_{Ii}(x,y) - \hat{I}_{Ti}(x,y)) && (9) \\ \mathcal{L}(V_I) &= -\sum_i \sum_{\forall x,y} (I_{Ii}(x,y) - \hat{I}_{Ti}(x,y)) && (10) \end{aligned}$$

d) **Noise estimation::** Once $(\Psi, \mathcal{P})$ are estimated from the data, the problem of estimating noise reduces to finding outliers that do not fit the model. Thus, for every frame i, we may classify pixels as noisy if they lie more than two standard deviations away from the predicted pixel color using the estimated values of $(\Psi, \mathcal{P})$.

e) **Outlier replacement::** Given the model $(\Psi, \mathcal{P})$, outliers can be replaced by projecting the model onto parts of the image labeled as outliers. This may be thought of as extrapolating the learnt model to predict missing data.

f) **Registration::** The problem of estimating $(\Psi, \mathcal{P})$ introduces registration into our framework. Registration of two frames of the input video $\mathcal{V}_{Ii}$ and $\mathcal{V}_{Ij}$ involves finding a correspondence between the frames, that best explains the visual data present in them. Thus, it may be represented by a function Φ, that takes $\mathcal{V}_{Ii}$ to $\mathcal{V}_{Ij}$. The function Φ is defined over $(\Psi, \mathcal{P})$.

$$\mathcal{V}_{Ii} = \Phi_{(i,j)}(\mathcal{V}_{Ij}) \quad (11)$$
$$\mathcal{V}_{Ij} = \Phi_{(j,i)}(\mathcal{V}_{Ii}) \quad (12)$$

The implicit assumption in this case, is that $\mathcal{V}_{Ii}$ and $\mathcal{V}_{Ij}$ have sufficient overlap of visual data, which is true for consecutive frames of a video. The main insight to note here, is that Φ is actually adequate to estimate and remove all the noise present in the input video. This is because we assume that (1) every part of the "true" video is present somewhere in the input and (2) that noise, unlike the rest of the visual data, is independently introduced in every frame of the video. In cases of occlusion, noise in frames are related, but not by the same function as the rest of the visual data. Thus, hypotheses of $(\Psi, \mathcal{P})$ may be replaced in equation (6) by $\hat{\Phi}$. Additionally, since we do not know which frame of the input sequence contains the "true" image or its parts, Eqn 7 has to be defined over all the images, for every image of V_I.

$$p(V_I|\hat{\Psi}, \hat{\mathcal{P}}) = \prod_j \prod_i p(I_{Ii}|\hat{I}_{Ij}) \quad (13)$$

In other words, every individual frame of the best hypothesis should not only explain the current image correctly, but should also be able to explain *all* the other images, when transferred through the registration function $\hat{\Phi}$. Thus, the overall log-likehood to be minimized becomes

$$\mathcal{L}(I_{Ii}) = -\sum_j \sum_{\forall x,y} (I_{Ii}(x,y) - \hat{\Phi}_{(i,j)} I_{Ij}(x,y)) \quad (14)$$
$$\mathcal{L}(V_I) = -\sum_i \sum_j \sum_{\forall x,y} (I_{Ii}(x,y) - \hat{\Phi}_{(j,i)} I_{Ii}(x,y)) \quad (15)$$

III. Results

In this section, we apply the theory developed earlier to two scenarios, which differ in their definitions of Ψ. This in turn defines noise, and hence determines what can be removed in each situation. The first scenario defines Ψ as a planar object and the second scenario defines Ψ as a bundle of rays. In each of these cases, we first compute SIFT [15] correspondences across frames, followed by a homography estimation based on the Gold Standard Algorithm [16]. Once pairwise homographies are computed, classification of each pixel of every image is done using homography based registration between frames. Computationally, the largest bottleneck is the feature extraction part which takes around 10 minutes for 500 frames of a video with resolution 640×480.

A. Scenario 1: Planar object

Figure 1 shows different frames of a planar object being observed from different views. Illumination artifacts are observed due to the specular nature of its texture. For the case of a planar scene, we define the registration function as a homography [16].

$$\Phi(i,j) = H(i,j) \quad (16)$$
$$\mathbf{x}_j = H(i,j)\mathbf{x}_i \quad (17)$$
$$\Phi(j,i) = H(j,i) = H^{-1}(i,j) \quad (18)$$

A homography between two frames may be computed from 4 accurate point correspondences. However, in the presence of noise, robust algorithms to estimate homographies exist [17]. Equation 17 describes the registration between points of the frames $\mathcal{V}_{Ii}$ and $\mathcal{V}_{Ij}$. Given pair-wise homographies, outliers are computed as points on the image whose colours are not consistent with other frames.

In practice, we do not evaluate over 255 grey levels, while computing Ψ to account for change in lighting conditions, and normalize each image before processing.

1) Experiment 1:: Figure 1 shows results of noise estimation and removal over a video sequence of a planar object. The camera moves arbitrarily over a board, while a light source produces a specularity on its surface. In this case, even the light source is in motion, though the board is stationary. However, our method equally applies to cases where any of the three objects are in motion. As can be seen, illumination artifacts are correctly identified (row 2 of figure), and replaced (row 3) to produce a video without the specular high light. The only input has been the nature of Φ. Everything else came out of the video automatically as the "noise".

Figure 3 shows results on the same scene, when different number of views are used to estimate and remove the noise in one particular frame of the sequence. In order to collect ground truth for this experiment, a frame without specularities was taken and view transferred by estimating homography using manually given correspondences. Intensity differences between this frame and the frames with noise removed, were used to derive the accuracy of our algorithm. As the graph shows, the accuracy reaches a saturation point after some threshold number of views.

B. Scenario 2: Rotating camera

When a camera pans, different views capture the same bundle of rays corresponding to every 3D point observed in more than one image. Different views taken with such a camera are thus related by the infinite homography [16]. Unlike the previous scenario, however, the definition of noise changes in case of panning cameras, though registration between frames is still represented as a homography. Since the infinite homography explains objects present at any arbitrary depth, dynamic objects present in the scene are estimated as outliers, and hence represent the noise.

Figure 4 shows frames of a video with people walking. Notice how in this definition the number of people make no difference to our algorithm. The second and third rows show the estimated outliers and the corresponding recovered images. Also shown is a mosaic generated from the recovered images, and the noise, which may have further applications along the lines of [19].

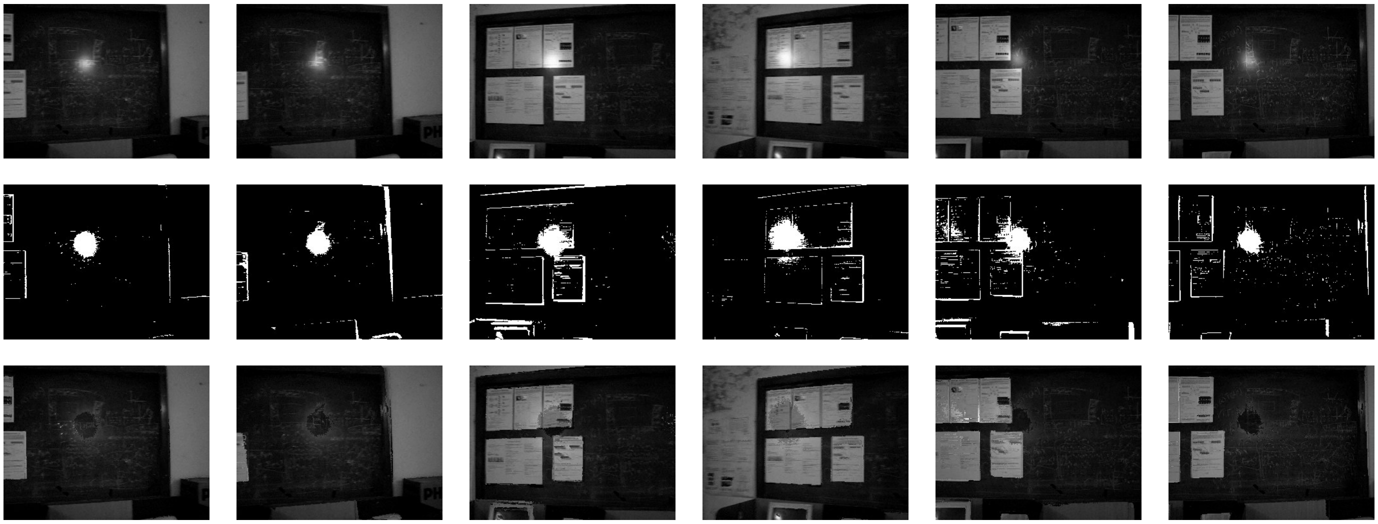

Fig. 1: Frames from a video sequence of a planar object with a specular surface. The first row shows frames 40, 100, 150, 200, 250 and 300 of a 306 frame sequence. The second row shows the outlier detection result. The final row shows reconstructed images after outlier removal.

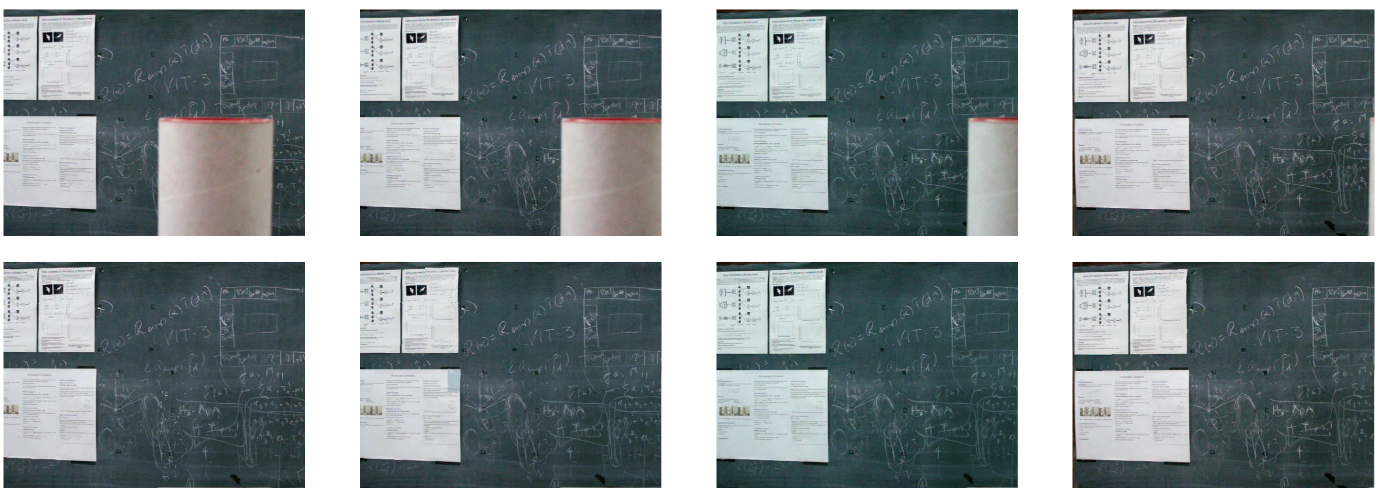

Fig. 2: Results for a 3D object observed by a moving camera with a planar background. In this case, the 3D object is estimated as noise, and removed. Frames 200, 250, 300, 350 of a 400 frame sequence are shown.

IV. Conclusion

In this paper, we presented an approach to automatically identify and remove "noise" from a video, based on a function. We then showed how the video completion problem can be posed as the removal of outliers given a proper function to be satisfied by the completed video. The identification, removal, and completion are performed automatically with no interactive inputs. We demonstarted its application on several representative videos.

The main drawback of the approach is the need for a function to be satisfied by the "true" video. The constraint function could be simple homographies as in the examples shown here. Complicated videos with independent motion of the camera and multiple objects may be hard to handle as the the underlying constraint functions are complex. Our algorithm, being unsupervised, also requires the true video's pixels to appear more frequently than the noisy ones. Artifacts can be seen in Figure 2, where sufficient frames were not available for identifying and replacing pixels. The best replacement pixel might not be easy to determine as multiple candidates obeying the model can exist. Figure 2 shows results for one such replacement strategy.

However, we believe automating the completion process is possible and advantageous in many situations. In addition, the machine learning framework allows for further extension to a variety of scenarios, involving dynamic 3D scenes with occlusions.

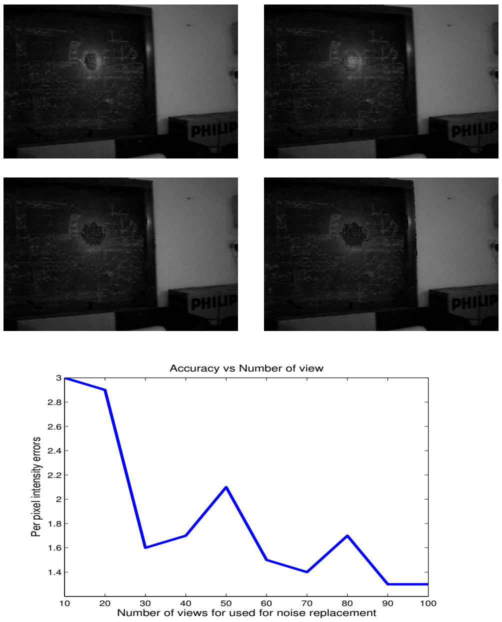

Fig. 3: Figures on the left show removal of lighting artifacts when 20, 50, 80 and 100 views are considered to estimate outliers. The graph on the right shows decreasing error when compared to ground truth. The x-axis represents number of views, and the y-axis, per-pixel intensity differences.

Fig. 4: Frames 70, 90 and 110 of a 200 frame video taken by a panning camera with multiple people walking. Notice how the last image has 4 people cluttered together. The third rows shows the estimated foreground.

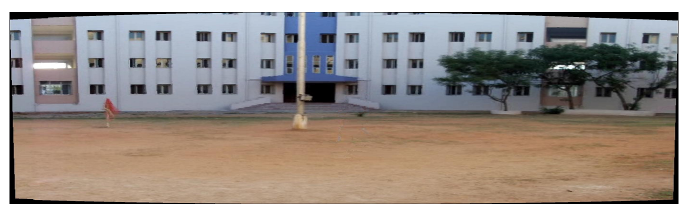

Fig. 5: A mosaic of the “true” video using the method of [18]

REFERENCES

[1] J. V. Candy, *Signal Processing: Model Based Approach*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[2] A. Kokaram, “On missing data treatment for degraded video and film archives: a survey and a new bayesian approach,” *IEEE Transactions on Image Processing*, pp. 397–415, 2004.

[3] A. Efros and T. Leung, “Texture synthesis by non-parametric sampling,” *International Journal of Computer Vision*, pp. 1033–1038, 1999.

[4] Y. Wexler, E. Shechtman, and M. Irani, “Space-time video completion,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 120–127, 2004.

[5] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, “Video inpainting for occluding and occluded objects,” *IEEE Conference on Image Processing*, pp. 69–72, 2005.

[6] V. Cheung, B. J. Frey, and N. Jojic, “Video epitomes,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 42–49, 2005.

[7] A. Kokaram, “Practical, unified, motion and missing data treatment in degraded video,” *Journal of Mathematical Imaging and Vision*, pp. 163–177, 2004.

[8] G. S. M. Bertalmio, A. L. Bertozzi, “Navier-stokes, fluid dynamics, and image and video inpainting,” *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 355–362, December 2001.

[9] J. Jia, T. P. Wu, Y. W. Tai, and C. K. Tang, “Video repairing: Inference of foreground and background under sever occlusion,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 364–371, 2004.

[10] M. Irani and S. Peleg, “Image sequence enhancement using mutliple motions analysis,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 282–287, 1992.

[11] S. P. Mallick, T. Zickler, P. N. Belhumeur, and D. J. Kriegman, “Specularity removal in images and videos: A pde based approach,” *European Conference on Computer Vision (ECCV)*, 2006.

[12] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg, “Dynamosaicing: Mosaicing of dynamic scenes,” *IEEE Pattern Analysis and Machine Intelligence*, 2007.

[13] D. Capel and A. Zisserman, “Super-resolution from multiple views using learnt image models,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[14] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[15] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal on Computer Vision*, pp. 91–110, 2004.

[16] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[17] K. ichi Kanatani and N. Ohta, “Accuracy bounds and optimal computation of homography for image mosaicing applications,” *International Confernce on Computer Vision*, pp. 91–110, 1999.

[18] M. Brown and D. Lowe, “Recognising panoramas,” in *International Conference on Computer Vision*, 2003, pp. 1218–1225.

[19] Y. Li, J. Sun, and H.-Y. Shum, “Video object cut and paste,” in *ACM SIGGRAPH*, 2005, pp. 595–600.