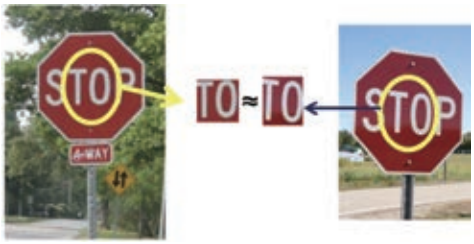
의미론적 연관성 분석을 위한 영상정합 동향

Ⅰ. 서 론

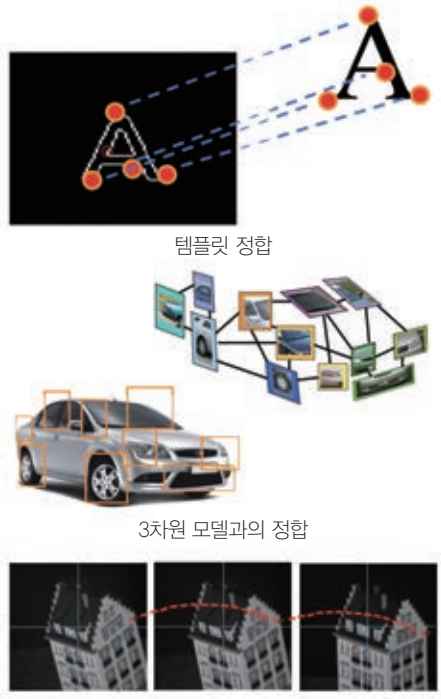
요즘은 핸드폰을 이용하면 손쉽게 파노라마 사진을 만들 수 있다. 사람들이 연속된 사진을 카메라로 촬영하면 알고리즘(Algorithm)이 자동으로 파노라마 사진으로 변환해준다. 어떻게 여러장의 사진을 이용해 파노라마 사진을 만들 수 있을지 생각해보자. 컴퓨터가 아니라 우리가 직접 파노라마 사진을 만드는 과정을 생각해보면 쉽게 답을 찾을 수 있다. 사람들마다 다양한 방법을 이용하겠지만, 공통적으로 사진들 사이에 존재하는 동일한 부분을 찾아 붙이고 중복되는 부분들을 제거할 것이다. 즉, 한장의 파노라마 사진으로 완성하기 이전에 사진들에 존재하는 동일한 부분을 찾을 것이다. 이처럼 두 영상 혹은 다중 영상 간에 대응되는 영역 혹은 대응점들을 찾는 과정을 영상 정합(Image correspondence)이라 한다(〈그림 1〉). 영상 정합은 컴퓨터비전 (Computer vision) 및 영상처리 (Image Processing) 분야의 핵심이다. 일례로 컴퓨터비전 분야의 저명한 교수인 카네기멜론 대학의 Takeo Kanade 교수는 컴퓨터비전 분야에서 가장 중요한 세가지 문제는 “Alignment, Alignment, Alignment” 라고 말하기까지 했다. 영상 정합은 파노라마 사진 이외에도 템플릿 정합, 3차원 모델과의 정합, 물체 추적 등에도 널리 이용되고 있다(〈그림 2〉). 앞으로 우리는 두 영상 사이에 대응점을 찾는 방법과 최근 연구 동향에 대해 살펴볼 것이다.

영상 정합은 정합의 밀도에 따라 성긴 정합(Sparse correspondence)과 밀집 정합(Dense correspondence)으로 나눌 수 있다. 성긴 정합은 영상의 특징점을 구한 뒤, 특징점들 사이에 대응되는 점을 찾는 과정이다. 이와 달리, 밀집 정합은 영상에 존재하는 모든 점(Pixel)들의 대응점을 찾는 과정이다. 대표적인 성긴 정합으로는 그래프 정합(Graph matching)이 있으며, 밀집 정합에는 스테레오 정합(Stereo

함 범 섭
연세대학교 전기전자공학부



〈그림 1〉 영상 정합의 예 (Image courtesy Derek Hoiem)



〈그림 2〉 영상 정합 응용 분야 (Image courtesy Silvio Savarese)

matching), 움직임 추정(Optical flow)이 있다. 스테레오 정합은 스테레오 카메라로 획득된 두 영상간의 정합점을 찾는 과정으로 이를 이용하면 깊이 정보를 획득 할 수 있다(〈그림 3〉). 본 기고에서는 밀집 정합, 나가아 밀집 시멘틱 정합 (Semantic correspondence)의 최근 연구 동향에 대해 알아보고자 한다.

Ⅱ. 의미론적 영상 정합 동향

1. 문제의 정의

시멘틱 정합의 연구 동향을 알아보기에 앞어 먼저 의미론적 정합, 혹은 시멘틱 정합의 정의에 대해 살펴보자.

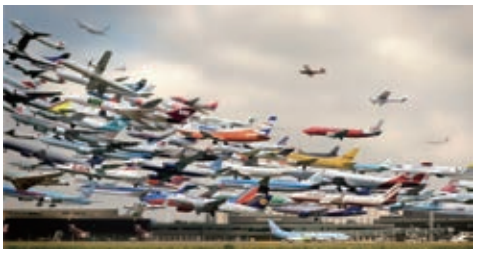
〈그림 3〉 스테레오 정합 (Image courtesy Sintel dataset)

〈그림 4〉 시멘틱 정합 (Image courtesy PASCAL VOC)

〈그림 4〉는 시멘틱 정합에서 궁극적으로 해결하고자하는 문제를 보여주고 있다. 기존의 정합(〈그림 3〉)과 달리 시멘틱 정합(〈그림 4〉)은 같은 물체 범주(Category)에 속해 있지만 서로 다른 물체에 대한 밀집 정합점을 찾는데 그 목적이 있다. 서로 다른 물체는 서로 다른 모양(Appearance), 크기 (Scale), 변형 (Deformation) 등을 가질 수 있다(〈그림 5〉). 따라서, 기존의 정합 방법을 그대로 이용하면 좋은 성능을 기대하기 어렵다.

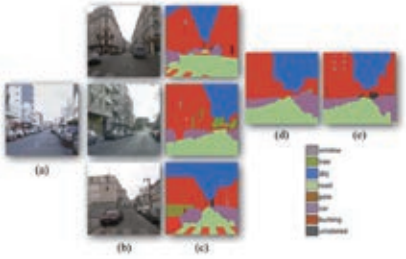
2. 응용 분야

시멘틱 정합을 이용하면 어떤 것들이 가능한지 살펴보자. 시멘틱 정합을 통해 물체 부분 혹은 물체 전체에 대한 정합을 구할 수 있으며, 나아가 물체간의 연관성을 분석할 수 있다. 그렇다면 물체간의 연관성이 의미하는 바



〈그림 5〉 시멘틱 정합의 어려움 (Image courtesy Andrea Vedaldi)

(a) 물체 검출[1]



(b) 레이블 전파[2]

(c) 영상 합성[3]

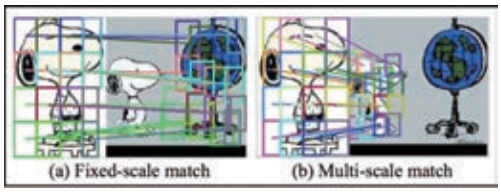
〈그림 6〉 시멘틱 정합의 응용 기술

는 무엇이며 이를 이용하면 무엇을 할 수 있을까? 사람이 어떤 작업을 할 때 물체 사이에 존재하는 연관성을 이용하는지 생각해보자. 인터넷 공간 상에는 수많은 영상들이 존재한다. 이 영상들 사이에 공통적으로 존재하는 물체 정보를 기계가 자동적으로 추출할 수 있을까? 공통적으로 존재한다는 것은 의미적으로 같은 물체를 찾는다는 것과 일치한다. 같은 물체를 찾기 위해서 사람은 무의식 중에 시멘틱 정합을 하고 있는 것이다. 즉, 수많은 영상을 사이에 시멘틱 정합을 수행하면 우리는 공통적으로 존재하는 물체를 추출해 낼 수 있다(〈그림 6(a)〉)[1]. 이 밖에도, 레이블링(Labeling)된 데이터베이스 상에서 질문(Query) 영상과 비슷한 영상을 찾은 후 레이블을 전파(Transfer)함으로써 질문 영상의 레이블을 자동으로 찾을 수 있으며(〈그림 6(b)〉)[2], 가상 영상을 합성할 수 있다(〈그림 6(c)〉)[3].

3. 방법론

시멘틱 정합은 일반적으로 〈그림 7〉과 같이 두 단계

〈그림 7〉 일반적인 시멘틱 정합 과정 (Image courtesy T. Taniai)



〈그림 8〉 SIFT Flow의 한계[14] (Image courtesy J. Kim)

를 요한다. 먼저 영상 내에 존재하는 모든 점들이 특징점이라고 가정한 뒤 이를 표현자(Feature descriptor)로 변환한다. 표현자는 영상 점들을 벡터 (Vector)로 변환하는 작업이다. 가장 쉽게는 컬러 영상의 RGB 값을 3차원 표현자로 이용할 수 있다. 하지만 앞서 살펴봤듯이 시멘틱 정합에서 다루는 물체는 모양, 크기 등이 변할 수 있으므로, 이러한 변형에 강인한 표현자를 사용한다. 대표적인 표현자로는 SIFT[4], HOG[5], DAISY[6] 등이 있으며, 최근에는 딥러닝(Deep learning)[7]을 이용한 CNN(Convolutional neural network) 기반의 표현자[8] 또한 이용되고 있다. 두 번째 단계는 두 영상에서 추출한 표현자 벡터들 사이의 거리를 측정함으로써 정합점을 구하는 과정이다. 일반적으로 하나의 표현자에 대해서 대응될 수 있는 모든 표현자와의 유클리디안 (Euclidean) 거리를 구하여, 그 중 가장 가까운 거리를 가지는 점을 대응점이라 한다.

최근에는 딥러닝을 이용하여 두 벡터 사이의 거리를 학습하는 연구도 활발히 연구 중이다[9-10].

4. 연구 동향

시멘틱 정합은 마이크로소프트(Microsoft)의 C. Liu에 의해 SIFT flow라는 방법으로 처음 제안되었다[11]. SIFT flow는 과거 움직임 추정에 널리 사용되는 에너지 함수를 차용하였으며[12], RGB값을 표현자로 이용하던 움직임 추정 방법과 달리 SIFT[4]를 표현자로 이용하였다. SIFT는 성긴 정합에 주로 사용되는 표현자로 물체의 크기와 방향

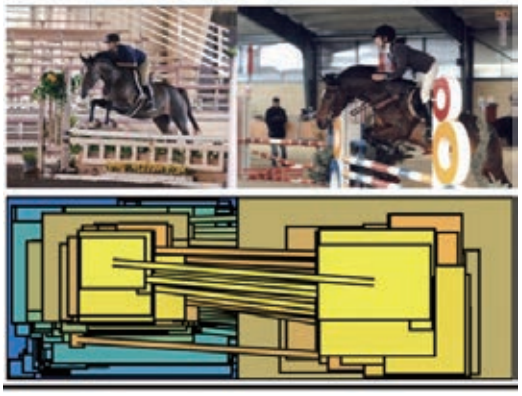
에 강인한 특징을 보인다. 따라서 C. Liu는 SIFT를 표현자로 이용하여 서로 다른 물체 사이의 정합점을 찾았다. SIFT flow는 〈그림 6〉에서 제시한 응용 기술 외에도 비디오에서 깊이를 추정하기 위해 사용되는 등 관련 분야에서 널리 이용되었다[13]. SIFT flow는 크기에 강인한 SIFT를 이용하였지만, 정작 SIFT에서 제공하는 크기 정보는 사용하지 않았다. 이는 정합하고자 하는 물체의 크기가 다를 때는 잘 동작하지 않는 한계를 보였다(〈그림 8〉). 이를 개선하기 위해 DSP 방법이 제안되었다[14].

DSP 방법은 에너지 함수 내에 크기에 관한 파라미터(Parameter)를 추가하여, 크기를 바꿔가면서 정합을 수행함으로써 SIFT flow의 문제를 해결하였다. 이와 함께 계층적인 최적화 방법을 제안하여 기존 SIFT flow가 가진 높은 연산량을 줄였다. DSP 방법은 물체 크기 변화에 강인하지만, 이 역시 기존의 SIFT 표현자에서 제공하는 크기 정보를 이용하지 않는 한계를 보였다. 이를 해결하기 위해 M. Tau는 영상 내에서 SIFT 표현자를 성기게 추출하고, 표현자에서 제공하는 크기 정보를 영상 내의 다른 곳으로 전파하는 방법을 고안했다[15]. 이를 이용하면 영상 전체에서 SIFT 표현자에서 제공하는 크기 정보를 이용해 정합을 수행할 수 있었다. M. Tau는 컬러 정보가 비슷하면 크기 정보가 비슷할 것이라는 가정을 이용하여 지역적인 방법(Local propagation)을 이용해 크기 정보를 전파했다.

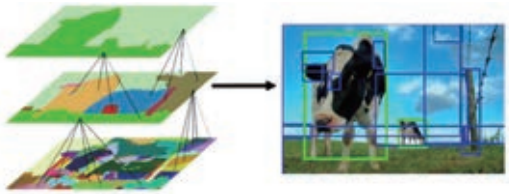
영상 내에 복잡한 배경이 있으면 물체들의 정합을 방해한다. 따라서 기존 시멘틱 정합 방법들은 배경 유무에 따라 많은 성능 차이를 보였다. 이를 해결하기 위해 최근 T. Taniai는 영상 세그멘테이션(Segmenation)을 함께 이용하는 시멘틱 정합 방법을 제안하였다(〈그림 9〉)[16]. 먼저 영상 내에서 전경과 배경을 분리하고 분리된 전경 사이의 정합점을 구하였다. 단순히 전경 배경을 분리하는 것이 아니라 여러 영상 내에 공통적으로 존재하는 전경을 분리함으로써 신뢰도를 높였다. T. Taniai는 이를 제귀적인 방법으로 풀어냈는데, 전경/배경 분리는 시멘틱 정합에 도움을 주고, 반대로 시멘틱 정합이 전경/배경 분리에 도움을 준다는 가정을 이용했다.

2016년에는 배경과 같은 클러터(Clutter)와 물체 크기 변형 모두에 강인한 Proposal flow라는 방법이 소개되었다(〈그림 10〉)[17]. 기존 정합 방법은 정합하려는 점에 대해 정해진 크기의 윈도우(Window)를 설정하여 이를 SIFT 등의 표현자로 바꾼 후 정합을 수행하였다. 이와 달리 Proposal flow는 object proposal을 정합의 기본 단위로 사용하였다. Object proposal은 기존 슬라이딩 윈도우 방법(Sliding window)을 대신하는 기법으로 영상 내에서 물체 혹은 물체 파트 같은 부분 만을 자동으로 추출해내는 기술이다[18, 19, 20](〈그림 11〉). 기존 슬라이딩 윈도우 방법이 물체가 될 모든 가능성을 염두해두고 영상 전체를 탐색하는 반면 object proposal을 이용하면 탐색 범위를 획기적으로 줄일 수 있다. 이러한 특성으로 인해 물체 인식[21], 물체 추적[22], 행동 인식[23], 시멘틱 세그멘테이션(Semantic segmentation) 등 여러 분야에 널

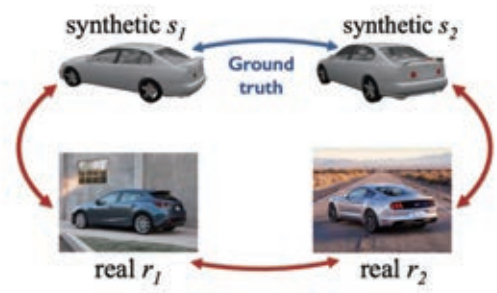
〈그림 9〉 Clutter에 강인한 시멘틱 정합 방법[16] (Image courtesy T. Taniai)



〈그림 10〉 Proposal Flow[17] (Image courtesy B. Ham)Taniai)



〈그림 11〉 Object proposal의 예[18] (Image courtesy J. R. Uijlings)



〈그림 12〉 Cycle consistency[24] (Image courtesy T. Zhou)

리 사용되고 있다. Proposal flow는 object proposal 간의 기하학적적 관계를 이용하여 시멘틱 정합의 정확도와 효율성을 끌어올렸다. 또한 성긴 object proposal 사이의 정합을 밀집 정합으로 보간하는 기술을 제안하여 기존 밀집 시멘틱 정합을 대체할 수 있는 가능성을 보여주었다.

시멘틱 정합 분야에서도 딥러닝을 이용하여 정합을 학습하는 방법이 소개되고 있다. T. Zhou는 3D 모델을 이용하여 시멘틱 정합을 학습할 수 있는 다량의 정답 데이터를 생성하였다. 3D 모델을 이용해 학습한 정합은 실제 영상에서 잘 동작하지 않기 때문에 실제 영상과 3D 모델에서 렌더링된 영상들 사이의 Cycle consistency를 이용하였다(〈그림 12〉). 즉, 3D 모델로부터 렌더링된 정답 정합이 실제 영상의 정합이 잘 되게 도와주는 방향으로 딥러닝 모델을 학습하였다. 이밖에도 시멘틱 정합을 위한 표현자를 학습하거나[25], 앞서 소개한 Proposal flow 방법을 딥러닝을 이용해 학습하는 방법[26]이 활발히 연구 중이다.

Ⅲ. 결 론

지금까지 시멘틱 정합의 정의, 응용 분야 및 연구 동향에 대해 살펴봤다. 시멘틱 정합은 컴퓨터 비전 및 영상처리 분야에서 그 중요성이 점점 대두되고 있으며, 최근 많은 연구 결과들이 소개되고 있다. 앞으로는 1) 시멘틱 정합을 위한 정답 데이터를 만드는 것과 나아가 2) 정답 없이 시멘틱 정합을 학습할 수 있는 방법에 관한 연구가 진행되리라 전망된다.

참 고 문 헌

[1] M.Rubinstein, A.Joulin, J.Kopf, and C.Liu,"Unsupervised joint object discovery and segmentation in internet images," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2013.

[2] C.Liu,J.Yuen,and A.Torralba, "Nonparametric scene parsing via label transfer," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 12, pp. 2368-2382, 2011.

[3] T. Hassner, "Viewing Real-World Faces in 3D", in Proc. IEEE International Conference on Computer Vision, 2013.

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

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2005.

[6] H. Yang, W.-Y. Lin, and J. Lu, "Daisy filter flow: A generalized discrete approach to dense correspondences," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2014.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Neural Information Processing Systems, 2012.

[8] J. L. Long, N. Zhang, and T. Darrell, "Do ConvNets learn correspondence?" in Proc. Neural Information Processing Systems, 2014.

[9] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2015.

[10] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2015.

[11] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," IEEE Transactions on Pattern Analysis and Machine Intelligence,

vol. 33, no. 5, pp. 978–994, 2011.

[12] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 3, pp. 500–513, 2011.

[13] K. Karsch, C. Liu, and S. Kang, "DepthTransfer: Depth extraction from video using non–parametric sampling," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2144–2158, 2014.

[14] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences, in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2013.

[15] M. Tau and T. Hassner, "Dense correspondences across scenes and scales," IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 38, no. 5, pp. 875–888, 2016.

[16] T. Taniai, S. N. Sinha, and Y. Sato, "Joint recovery of dense correspondence and cosegmentation in two images," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2016.

[17] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2016.

[18] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," International Journal of Computer Vision, vol. 104, no. 2, pp. 154–171, 2013.

[19] P. Arbelaez, J. Pont–Tuset, J. Barron, F. Marques, and J. Malik, "Multi–scale combinatorial grouping," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2014.

[20] J. Hosang, R. Benenson, P. Dolla'r, and B. Schiele, "What makes for effective detection proposals?" IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malic, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2014.

[22] G. Zhu, F. Porikli, and H. Li, "Beyond local search: Tracking objects everywhere with instance–specific proposals," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2016.

[23] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r*CNN," in Proc. IEEE International Conference on Computer Vision, 2015.

[24] T. Zhou, P. Krä̧henbü̧hl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D–guided cycle consistency," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2016.

[25] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "FCSS: Fully convolutional self–similarity for dense semantic correspondence," in Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2017.

[26] K. Han, R. S. Rezende, B. Ham, K.–Y. K. Wong, M. Cho, C. Schmid, and J. Ponce, "SCNet: Learning semantic correspondence," CoRR, vol. abs/1705.04043, 2017.

함범섭

- 2013년 8월 박사학위, 전기전자공학, 연세대학교
- 2008년 2월 학사학위, 전기전자공학, 연세대학교
- 2016년 9월~현재 연세대학교 전기전자 공학부 조교수
- 2014년 5월~2016년 7월 INRIA & École Normal Supérieure, 박사후 연구원
- 2013년 9월~2014년 4월 연세대학교 전기전자 공학부, 박사후 연구원

〈관심분야〉
컴퓨터비전, 영상처리, 기계학습