Survey on Vision-Based Path Prediction

Tsubasa Hirakawa[1], Takayoshi Yamashita[1], Toru Tamaki[2(✉)], and Hironobu Fujiyoshi[1]

[1] Chubu University, Aichi 487-0027, Japan
hirakawa@mprg.cs.chubu.ac.jp, {yamashita,hf}@cs.chubu.ac.jp
[2] Hiroshima University, Hiroshima 739-8527, Japan
tamaki@hiroshima-u.ac.jp

Abstract. Path prediction is a fundamental task for estimating how pedestrians or vehicles are going to move in a scene. Because path prediction as a task of computer vision uses video as input, various information used for prediction, such as the environment surrounding the target and the internal state of the target, need to be estimated from the video in addition to predicting paths. Many prediction approaches that include understanding the environment and the internal state have been proposed. In this survey, we systematically summarize methods of path prediction that take video as input and extract features from the video. Moreover, we introduce datasets used to evaluate path prediction methods quantitatively.

Keywords: Path prediction · Trajectory · Pedestrian · Survey
Datasets

1 Introduction

Path prediction is the task of estimating the path, or trajectory, along which a target (e.g., a pedestrian or vehicle) will move. Predicting paths from video is an important task receiving much attention as it is expected to have many potential applications, such as surveillance camera analysis, self-driving cars, and autonomous robot navigation.

Path prediction has to estimate much more information—such as information of the surrounding environment, moving direction, and status of prediction targets—than other simple image recognition tasks. As a result, prediction methods are often built on top of other computer vision tasks, such as pedestrian detection [1,2], pedestrian attribute recognition [3], and semantic segmentation [4]. Moreover, in the prediction task, future observations of predicted paths are not available. In tasks of pedestrian detection and tracking, observations from the past to the present are used to locate and track the target in the current frame of the video. In contrast, the prediction task localizes and predicts the locations of the target in future frames of the video, using observations made until the present time and prior information on the surrounding environment and knowledge of the target motion.

N. Streitz and S. Konomi (Eds.): DAPI 2018, LNCS 10922, pp. 48–64, 2018.
https://doi.org/10.1007/978-3-319-91131-1_4

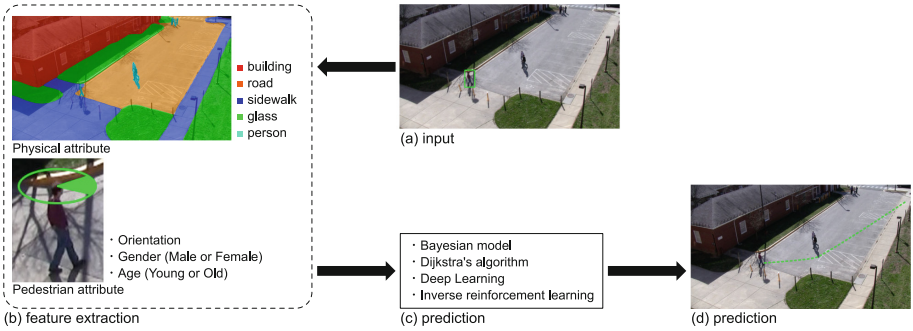


Fig. 1. Overview of path prediction, modified from [6].

Path prediction has been studied for decades in the field of robotics. At stations and airports, robots need to move without interfering with the many people present [5] and to plan a path of efficient motion in the environment. Path prediction is necessary to achieve such tasks. However, in addition to information from cameras, robots are able to use information from many types of sensor, such as a LIDAR sensor, to obtain the three-dimensional (3D) geometry of the scene. The environment in which the robot can move around is sometimes explicitly given as an environment map. The present survey is of path prediction methods involving video only as a computer vision task.

There is an alternative task called early recognition, which predicts future human behaviors in video. This task predicts future actions in the video but is excluded from the survey because the predicted categories are discrete whereas predicted paths are sequences of continuous locations.

As the task of path prediction in the field of computer vision is difficult and challenging, a number of various methods have been proposed. A common approach is shown in Fig. 1. As input, a video (or a frame of video) is given in addition to the location of the target in the current frame or a sequence of locations over the past frames of several seconds. Features useful for prediction are then extracted from the video (or frames) to predict the path in future frames. There are two important parts to the overview of Fig. 1: (b) feature extraction where many features are extracted to understand the environment and target; and (c) path prediction where a variety of methods are proposed, categorized into four types.

In this paper, we survey path prediction methods taking video as input and systematically summarize feature extraction and prediction approaches and datasets used for evaluation. We explain feature extraction methods in Sect. 2 and categorize prediction methods in Sect. 3. In Sect. 4, we review datasets used in evaluating the performance of path prediction. We conclude the survey in Sect. 5.

Table 1. Categories of feature extraction for path prediction.

Feature	Types	Methods
Environment	Scene label	Stacked hierarchical labeling [7]
		Superpixel-based MRF [8]
		Fully convolutional networks [9,10]
	Cost	Bag of visual words
		Spatial matching network [11]
	Global scene feature	Pre-trained AlexNet [12]
		Siamese network [13]
Target	Location	HOG + SVM detector [14]
	Direction	Bayesian orientation estimation [15]
		Orientation network [11]
	Attribute	AlexNet-based multi-task learning [16]
	Feature vector	Mid-level patch features [17]

2 Feature Extraction from a Video

This section introduces methods of feature extraction from video for path prediction. The path that the pedestrian takes is implicitly affected by many factors of the surrounding environment and the status of the pedestrian his self or herself. The performance of path prediction is expected to be improve when using information that largely determines how the pedestrian decides the way to go. Given the video, such information is extracted prior to the prediction. Table 1 presents information extracted from video for path prediction. Such information can be broadly categorized into that of (1) the environment and (2) the target.

2.1 Environmental Features

The pedestrian decides the way and walks along a path while being affected by the surrounding environment. For example, we usually walk along the sidewalk while avoiding obstacles on the way (e.g., parked cars and trash cans) and drive a car on the roadway as is common social practice. The movement of the target is dynamically affected by the environment, and environmental features are therefore extracted from the video.

Semantic segmentation [18–21] is a task of assigning an object class to each pixel, which is the most common task in understanding the environment in the field of computer vision. Semantic segmentation can be conducted to estimate where obstacles exist in the scene and where there are regions available for walking. Kitani et al. [18] assumed that pedestrian paths are mainly affected by the physical environment, such as sidewalks, roadways, flower beds, and buildings, and predicted posterior probabilities of each label using hierarchical segmentation [7] as shown in Fig. 2. These probabilities are used as feature vectors to

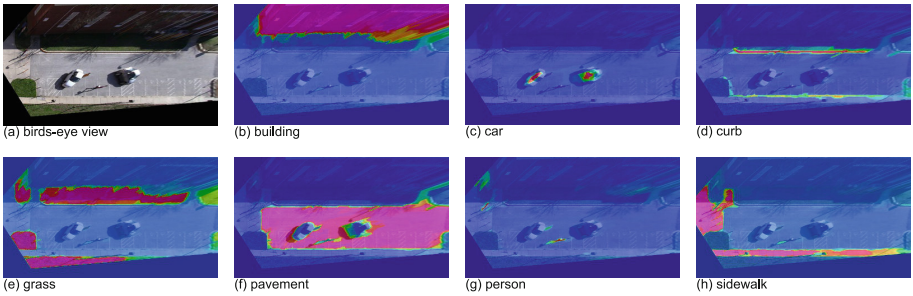


Fig. 2. Examples of environmental attributes [18]

form scene feature maps, which are used for path prediction. Rehder et al. [20] used segmentation results obtained using a fully convolutional network [9,10] for prediction.

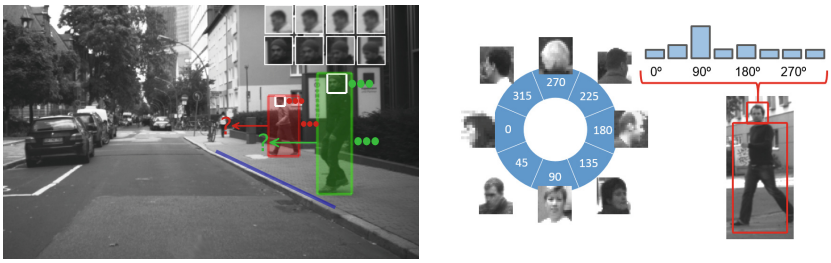
Alternative approaches do not explicitly use environmental features affecting paths but implicitly represent probabilities of paths as cost (or reward) functions [11,22]. These methods create cost maps of the entire scene from cost functions independently estimated for each superpixel. Walker et al. [22] searched for patches that have similar texture from training samples using a nearest-neighbor approach, and assigned the costs of the training samples to superpixels to generate cost maps of the scene. Huang et al. [11] proposed a convolutional neural network (CNN) called the spatial matching network, which estimates rewards of local regions by comparing similarity between the patch of the target and surrounding superpixel patches.

Yet another approach represents the scene as a single feature vector, whereas the above approaches extract local features from superpixels. Assuming that similar scenes prompt similar paths, this approach retrieves similar scenes in a training dataset with feature vectors to predict paths using the paths of the retrieved scenes. To this end, CNNs are usually used to efficiently extract scene feature vectors because of the recent success of deep learning architectures. In predicting paths in first-person video, Park et al. [23] used AlexNet [12] to extract features when retrieving scenes, and transferred paths of the retrieved scenes for prediction. Su et al. [24] used an AlexNet-based Siamese network [13] to retrieve features.

2.2 Target Features

While environmental features strongly affect the target in terms of the path decision, internal factors of the target are also important. Specifically, attributes of the target, such as age, gender, and internal demand, affect the path decision. We herein introduce methods for extracting target features.

The most common target feature is the orientation of the target [11,16,25] because the estimated orientation can be used to predict in which direction the target is going. In other words, the orientation constrains the moving direction



(a) (b)

Fig. 3. Estimation of head orientation [25]. (a) Detection of heads and bodies of pedestrians. (b) Estimation of the orientation of the head in eight directions.

of the target and thus reduces errors of prediction. Kooij et al. [25] detected pedestrians employing a histogram of oriented gradients (HOG) and support vector machine (SVM) [14] and estimated the head orientation [15] to predict the path of a pedestrian in front of a car on which a camera was mounted, focusing on whether the pedestrian will stop before stepping forward onto the roadway as shown in Fig. 3. If the head faces the camera, then the pedestrian is assumed to notice the car and is predicted to slow down or stop before the roadway.

Physical attributes, such as age and gender, are also important to prediction. When walking in places where there are a number of people, pedestrians take actions to avoid colliding with each other. Aspects of such avoidance—when and where pedestrians start to avoid others—are different for pedestrians of different age and gender; e.g., a younger person walks faster and responds more rapidly to others than senior people. Wei et al. [16] used AlexNet to estimate the orientation, age, and gender of pedestrians as multi-task learning. Estimated attributes are used in deciding the walking speed of pedestrians.

Walker et al. [22] proposed unsupervised path prediction by extracting mid-level feature vectors directly from patches containing the target, instead of direct attributes.

3 Prediction Methods

Path prediction follows feature extraction from video. Table 2 summarizes methods of prediction, categorized according to their approach. This section describes each category and its properties.

3.1 Bayesian Models

The first approach uses online Bayes filters, such as Kalman filters (KFs) and particle filters, and infers the model to predict paths. Such modeling introduces internal states and observations as variables, and defines probabilistic models by

Table 2. Categories of path prediction methods

Category	Paper	Year	Method	Scene	Input	Output	Feature	
							Env.	Target
Bayesian	[26]	2013	KF	Car	Coord.	Coord.		
	[25]	2014	DBN	Car	Video	Coord.	✓	✓
	[19]	2016	DBN	Top view	Video	Coord.	✓	
Energy minimization	[27]	2013	Dijkstra	Top view	Video	Distribution	✓	
	[22]	2014	Dijkstra	Surveillance	Video	Distribution	✓	
	[11]	2016	Dijkstra	Surveillance	Image	Distribution	✓	✓
DL	[28]	2016	CNN	Surveillance	Coord.	Coord.		
	[29]	2016	LSTM	Top view	Coord.	Coord.		
	[30]	2017	LSTM	Top view	Coord.	Coord.		
	[31]	2017	LSTM	Top view	Coord.	Coord.		
	[21]	2017	RNN Enc.-Dec	Car	Video	Coord.	✓	
IRL	[18]	2012	IRL	Top view	Video	Distribution	✓	
	[32]	2016	IRL	Top view	Video	Distribution	✓	
	[33]	2016	IRL	First person	Video	Distribution		✓
	[34]	2017	IRL	First person	Video	Distribution		✓
	[16]	2017	IRL	Surveillance	Image	Distribution		✓
	[20]	2017	IRL	Car	Video	Distribution	✓	
Others	[35]	2014	Optical flow	Car	Video	Coord.		
	[36]	2015	Markov process	Car	Video	Coord.	✓	
	[23]	2016	Data driven	First person	Video	Coord.	✓	
	[24]	2017	Data driven	First person	Video	Coord.	✓	
	[37]	2011	Social force	Top view	Video	Coord.		
	[38]	2016	Social force	Top view	Video	Coord.		

assuming that the observations are the internal states contaminated by noise. This approach iterates the prediction step that computes the current internal states from the previous states, and the update step that updates the current states with the observations. In a common setup, internals states are actual coordinates of pedestrians, and observations are coordinates obtained by pedestrian detection. This is person tracking if we apply the approach to track from the past to present, and path prediction if we only repeat the prediction step to obtain the sequence of coordinates of the pedestrian, without the update step; i.e., there are no future observations.

Schneider et al. [26] used the extended KF to update the internal state of the pedestrian in front of a car. This was an early work of path prediction and showed what kind of primitive information (e.g., the walking speed and acceleration) is useful for path prediction.

Instead of using online Bayes filters, some works have used a dynamic Bayesian network (DBN) [19,25]. Kooij et al. [25] considered a more restricted case; estimating if the pedestrian will walk across a roadway in front of a car on

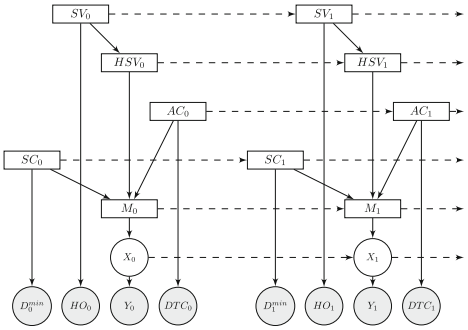


Fig. 4. Graphical model of a DBN with an SLDS [25]

which a camera is mounted. They defined a DBN model with a switching linear dynamical system (SLDS) that is shown in Fig. 4 and that uses features extracted from the movie, such as the pedestrian's head orientation, distance to the car, and distance between the pedestrian and roadway. This method performs better than using coordinates of pedestrian detection only.

3.2 Energy Minimization

The Bayes approach described above is on-line in which it estimates the coordinates of the pedestrian frame by frame in the video. Another (off-line or batch) approach is an energy minimization approach that estimates the entire sequence of coordinates at the same time. This approach constructs a two-dimensional grid graph of the scene and assigns costs for moving to edges in the graph, and then finds the combination of edges that gives the minimum energy. This is formulated as a shortest path problem solved employing the Dijkstra method. The prediction accuracy is therefore largely affected by how the cost is defined.

Huang et al. [11] proposed a path prediction method using a single image. First, a patch containing the target is extracted to estimate the orientation of the target. Next, the cost for moving across the location of the patch is estimated by comparing the texture of surrounding patches. In addition to this cost, the estimated orientation of the target is used as a constraint and added to the edge weights. Walker et al. [22] compared the texture of superpixels using patches along the path that the target traced without involving any training procedure.

Appearance information (texture) of the scene can be used to define the cost function, but objects in the scene can also be used. Xie et al. [27] assumed that pedestrians have decided their goal (e.g., a food trunk) according to their potential demands (hunger), and defined cost maps where the pedestrians are attracted to objects in the scene.

3.3 Deep Learning

Deep learning methods such as those involving the CNN and long short-term memory (LSTM) have been used for path prediction since the emergence of deep

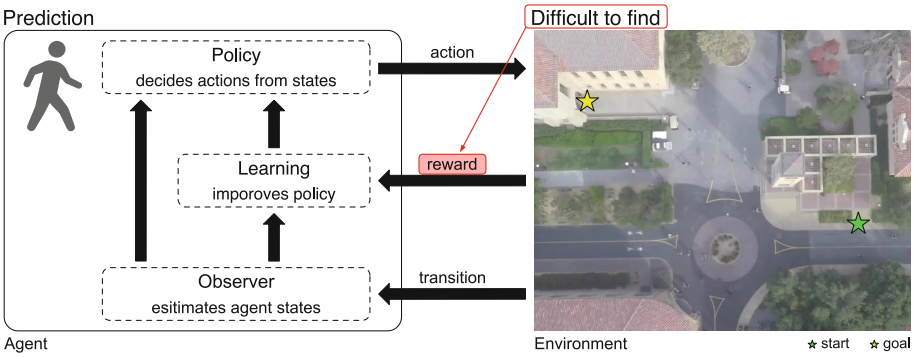


Fig. 5. Overview of RL, modified from [38].

learning frameworks. Methods of this type take as input the series of coordinates of the target over the last several frames, and produce a series of target coordinates in several successive frames. Feature extraction, described in the last section, is not explicitly performed as feature extraction and prediction are not explicitly separated in deep learning models.

Several methods thus use LSTM to deal with paths, which are sequences of two-dimensional coordinates, have been proposed. Alahi et al. [29] proposed the social-pooling (S-pooling) layer for avoiding collisions between pedestrians. A pedestrian is represented by LSTM, and hidden layer outputs of LSTMs of other people are connected to the S-pooling layer of the pedestrian. This layer allows the LSTM of the pedestrian to represent the spatial relationship with nearby people (e.g., the distance to each other), and thus predict the path avoiding collision.

LSTM has a limitation of long-term memory; i.e., paths in the distant future are difficult to predict. Fernando et al. [31] assumed the necessity of more elaborate long-term memory, and proposed the tree memory network that hierarchically selects useful information of the past stored in memory cells and performs better than other LSTM models.

Besides LSTM, the CNN is also used to directly make predictions. Yi et al. [28] proposed the behavior-CNN that predicts the future path from the past path. This method first creates three-dimensional sparse data whose channels store the pedestrian two-dimensional coordinates of the last several frames. The sparse 3D data are encoded using convolution and pooling layers and then decoded using deconvolution layers. They also added location bias maps to each channel of encoded information to account for different behaviors at different locations in the scene, such as the locations of entrances and obstacles.

3.4 Inverse Reinforcement Learning

The three approaches above are examples of supervised or unsupervised learning, while the approach presented here is an example of reinforcement learning (RL).

RL learns a policy to decide actions to be taken by an agent under the current status in an environment. RL is usually defined as a Markov decision process that learns the optimal policy to allow the agent to take the best actions maximizing the reward. Figure 5 shows that an agent of RL is the target of prediction, an environment is the scene given as video, a status is the pedestrian location, and an action is the movement of the pedestrian.

RL needs to define the reward of the action of moving from one state to another, which indicates how good the action taken by the agent is. However, it is difficult to explicitly define the reward function for practical problems such as the path prediction task. This problem is called the reward design problem, and inverse reinforcement learning (IRL) is one approach taken to solve the problem. IRL estimates rewards that reproduce optimal sequences of actions, and decides actions of the agent in the test phase with the estimated reward so that the agent can take similar actions.

IRL has been used to learn and control the optimal motion of robots [5]. Kitani et al. [18] first introduced IRL to vision-based path prediction. Instead of estimating target locations, they estimated actions that the agent may take at a certain time or location, and predicted possible paths by sequentially applying the estimated actions to the current target location. This task is therefore called activity forecasting, in contrast to path prediction that directly estimates locations of the target in the future. Activity forecasting is a much more complex and challenging task than path prediction while it has great potential in terms of having a variety of predictions adapted to each possible application.

Kitani et al. [18] assumed that the physical attributes of a scene strongly affect pedestrian paths, and used scene attributes estimated by semantic segmentation as feature maps. Rewards of each scene attribute are defined by the inner product of the feature maps and weight vectors, and the optimal weights are estimated from training data. For prediction, a sequence of actions that arrive at the predefined goal is generated by giving the goal and the current location of the target pedestrian. Lee et al. [32] used a similar approach to predict paths of football players in a game video. Wei et al. [16] introduced a game theory called fictitious play to predict paths of multiple pedestrians who arrive at a goal while avoiding collisions between pedestrians.

Without any predefined goals, Rehder et al. [20] proposed the destination network to estimate the goal of the target using the last several frames. The estimated goal and the environmental attributes obtained using a fully convolutional network were used to predict pedestrian paths.

For first-person vision, Bokhari et al. [33] used objects held by a person and the object states to predict goals in the future. While this work considered a limited scene (e.g., a kitchen), Rhinehart et al. [34] dealt with wider areas, such as a home including a kitchen, bathroom, and living room.

3.5 Other Approaches

Most prediction methods can be categorized into one of the four approaches described above, but there are other approaches.

Fig. 6. Prediction from first-person videos; (left) [23], (right) [24].

The social force model [39] assumes energy called a "social force" that acts between pedestrians and objects in the scene, and generates pedestrian movement through interaction via the force. Yamaguchi et al. [37] proposed a model with additional states, such as the preferences of pedestrians, walking speeds, goals, and the existence of other people walking together. This work was motivated by a desire to improve the accuracy of pedestrian tracking, but performed path prediction to evaluate the proposed model. Robicquet et al. [38] proposed social forces of multiple classes for avoiding collisions. They estimated "social sensitivity features" using the distances between other people, and applied K-means clustering of the features to get several clusters of avoidance behaviors. The cluster of the target behavior of avoidance was estimated using the target feature, and paths of the cluster were then projected back to the scene for prediction.

Optical flow extracted from car-mounted cameras was used by Keller et al. [35] to predict pedestrian paths. They used optical flows over the last several frames and computed orientation histograms as motion features of pedestrians. The sequence of histograms was used to retrieve similar scenes in the training set, and paths of the retrieved scenes were then mapped back to the scene for prediction.

The use of the Markov process framework was proposed by Rehder et al. [36]. They used normal and von-Mises distributions to represent the state (location) and speed of the pedestrian, and sequentially estimated the state by taking products of these distributions at each time step for prediction. To improve accuracy, the goal of the pedestrian was estimated from environmental attributes to constrain the direction of motion.

The retrieval-based approach shown in Fig. 6 was proposed by Park et al. [23] to predict the future path in a video showing the first-person view. They first extracted scene features using AlexNet and then found similar scenes in the training set by comparing extracted features. Paths of retrieved training samples were mapped onto the video. They predicted paths even in scenes with occlusions by estimating regions behind occluding objects, such as walls and obstacles. Su et al. [24] extended this work to the prediction of multiple basketball players

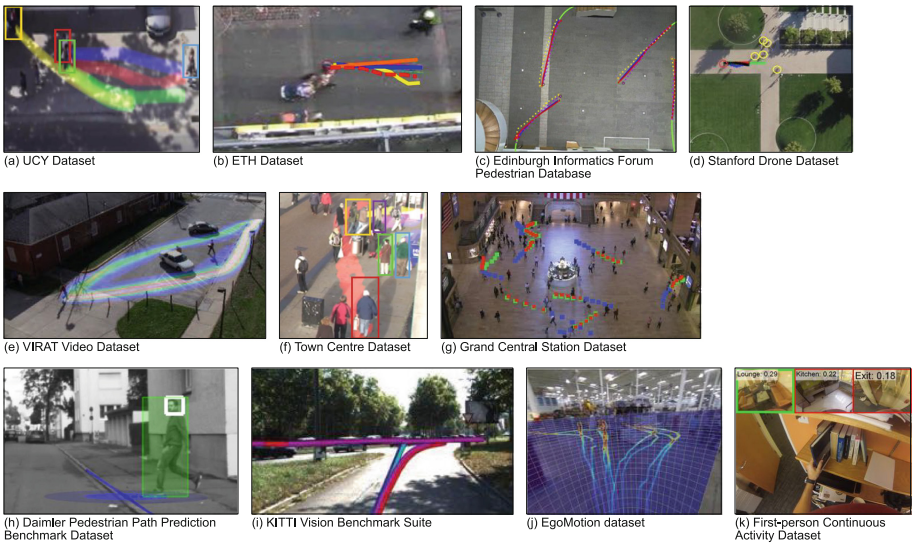


Fig. 7. Datasets and results of prediction, taken and modified from [16,18,21,23,25,28,29,31,34,38].

in a game scene. In one first-person video, they estimated the region of "joint attention" to which multiple players commonly paid attention. Multiple paths were predicted by selecting the optimal path of each player and by minimizing an objective function defined by the estimated joint attention region, locations of players, and paths projected back to the scene.

4 Datasets

This section briefly introduces datasets used to evaluate path prediction methods. Various datasets have been used as shown in Table 3 and Fig. 7. The diversity of datasets is due to the difficulty of using a single universal dataset for many different conditions, e.g., different numbers of scenes and paths needed for learning and different types of scenes. We therefore categorize datasets into four categories in terms of the viewpoint of the camera.

4.1 Videos of Entire Scenes

The most commonly used type of dataset is video that captures the entire scene taken by a wide-angle camera (for surveillance) at stations and market places. These datasets are usually used to evaluate pedestrian tracking methods; however, they are also used in evaluating path prediction because sequences of pedestrian locations are given as the ground truth.

Table 3. Comparison of datasets

	Year	URL	#People	Viewpoint	#Scenes	Other targets	Additional information
UCY [40]	2007	1	786	Top view	3	–	–
ETH [41]	2009	2	750	Top view	2	–	–
Edinburagh informatics forum [42]	2009	3	95,998	Top view	1	–	–
Stanford drone [38]	2016	4	11,216	Top view	8	Bikers, skateboarders, cars, buses, golf carts	–
VIRAT [6]	2011	5	4021	Surveillance	11	Car, bike	Object coordinates, activity category
Town centre [43]	2011	6	230	Surveillance	1	–	Head coordinates
Grand central station [44]	2015	7	12,600	Surveillance	1	–	–
Daimler [26]	2013	8	68	Car	–	–	Stereo camera
KITTI [45]	2012	9	6336	Car	–	Car	Stereo camera, LIDAR, Map
EgoMotion [23]	2016	–	–	First person	26	–	Stereo
First-person continuous activity [34]	2017	–	–	First person	17	–	Object information

1: https://graphics.cs.ucy.ac.cy/research/downloads/crowd-data
2: http://www.vision.ee.ethz.ch/en/datasets/
3: http://homepages.inf.ed.ac.uk/rbf/FORUMTRACKING/
4: http://cvgl.stanford.edu/projects/uav_data/
5: http://www.viratdata.org/
6: http://www.robots.ox.ac.uk/~lav/Papers/benfold_reid_cvpr2011/benfold_reid_cvpr2011.html
7: http://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html
8: http://www.gavrila.net/Datasets/Daimler_Pedestrian_Benchmark_D/daimler_pedestrian_benchmark_d.html
9: http://www.cvlibs.net/datasets/kitti/

Top view

The UCY Dataset [40] and ETH Dataset [41] contain videos of pedestrians walking along streets where no other moving objects exist, which is a relatively simple situation compared with situations of other datasets. The Edinburgh Informatics Forum Pedestrian Database [42] consists of videos of pedestrians walking at the campus of the University of Edinburgh taken by a fixed camera. This dataset is large and has more than 90,000 paths.

The above datasets are constructed for pedestrian tracking and crowd behavior analysis, while the Stanford Drone Dataset [38] focuses on path prediction. This dataset has videos taken by drones flying at eight sites of Stanford University, and provides annotations of moving objects, such as cyclists, skateboarders, and cars, as well as pedestrians.

Surveillance

Videos in the datasets described above are taken from a top view, while videos in the datasets shown in Fig. 7(e, f) are taken from a bird's eye view; i.e., the videos are taken by surveillance cameras looking downward at an angle. The physical attributes of pedestrians are observable in these videos and can be used for prediction. The VIRAT Video Dataset [6] contains videos taken by surveillance cameras at parking lots, and provides the locations of pedestrians, cars, and objects in the scene and labels of activities, such as getting into a car and opening a trunk. It contains 11 scenes, which is the largest number of scenes among the datasets of surveillance cameras in Table 3. The Town Centre Dataset [43] contains videos of pedestrians and provides bounding boxes of each pedestrian as well as labels of head locations of pedestrians.

The Grand Central Station Dataset [44] contains videos taken by a fixed camera mounted at a station, as shown in Fig. 7(g). It has a single scene but is complex owing to the many people appearing in and disappearing from the scene because the motivation is to analyze the behaviors of many pedestrians.

4.2 Car-Mounted Cameras

Datasets of videos taken by cameras mounted on vehicles are used because path prediction is studied with the aim to develop automated driving. In this case, cameras are mounted in the front of the car to look forward, and the main objective is to predict paths of pedestrians in front of the car.

The Daimler Pedestrian Path Prediction Benchmark Dataset [26] consists of videos taken by car-mounted cameras. There are four classes of cases, including cases that the pedestrian walks across the roadway and cases that the pedestrian stops walking to avoid an accident. In addition to the videos themselves, depth information is available as the videos are taken by stereo cameras. There are relatively few pedestrians; however, the dataset contains videos that are rare in other datasets, such as videos of pedestrians crossing in front of moving cars.

The KITTI Vision Benchmark Suite [45] was constructed for the Intelligent Transport System, and is used for various evaluations such as those of the detection of pedestrians, vehicles, and white lines on the road. It contains not only RGB images but also stereo images, LIDER 3D data, GPS locations, and street maps, and it is therefore useful for path prediction that uses rich information to understand the environment.

4.3 First-Person View

Unlike videos of entire scenes and taken by car-mounted cameras for predicting paths of targets in the scene, videos taken from the first-person view are used to

predict the path of the person taking the video. Park et al. [23] used first-person videos taken by wearable cameras moving through indoor and outdoor environments of 26 different scenes, such as on a street and inside a store. Rhinehart et al. [34] collected first-person videos taken by a person walking around office environments and assumed that an object held by the person (e.g., a mug or towel) indicate where the person is going (e.g., the kitchen or bathroom).

5 Conclusions

We reviewed vision-based path prediction methods and common datasets. We first categorized feature extraction methods of features used for prediction attributed to the environment or target appearance and dynamics. We then grouped prediction methods according to the approach taken. Bayesian methods define probabilistic models of the path and sequentially estimate internal states. Energy minimization methods define a two-dimensional grid graph by computing possibilities of pedestrians to move in each local region, and then solve the shortest-path problem. Deep learning methods take a series of locations of the target over the past several seconds and output a series of future locations. IRL uses the policy and reward estimated from training samples and then selects actions iteratively to produce a future path. These approaches are of course not exclusive and often used in combination [21]. Finally, we summarized datasets used in evaluating prediction methods. Some datasets are used for pedestrian detection and tracking, while others are used for path prediction.

Acknowledgments. This work was supported in part by JSPS KAKENHI under grant number JP16H06540.

References

1. Weinland, D., Ronfard, R., Boyer, E.: A survey of vision-based methods for action representation, segmentation and recognition. Comput. Vis. Image Underst. **115**(2), 224–241 (2011)
2. Benenson, R., Omran, M., Hosang, J., Schiele, B.: Ten years of pedestrian detection, what have we learned? In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 613–627. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_47
3. Deng, Y., Luo, P., Loy, C.C., Tang, X.: Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia, MM 2014, pp. 789–792. ACM, New York (2014)
4. Zhu, H., Meng, F., Cai, J., Lu, S.: Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation. J. Vis. Commun. Image Represent. **34**, 12–27 (2016)
5. Ziebart, B.D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J.A., Hebert, M., Dey, A.K., Srinivasa, S.: Planning-based prediction for pedestrians. In: International Conference on Intelligent Robots and Systems, pp. 3931–3936, October 2009

6. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J.K., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: Computer Vision and Pattern Recognition, pp. 3153–3160, June 2011
7. Munoz, D., Bagnell, J.A., Hebert, M.: Stacked hierarchical labeling. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 57–70. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15567-3_5
8. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. In: Computer Vision and Pattern Recognition, pp. 3294–3301 (2014)
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
10. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
11. Huang, S., Li, X., Zhang, Z., He, Z., Wu, F., Liu, W., Tang, J., Zhuang, Y.: Deep learning driven visual path prediction from a single image. IEEE Trans. Image Process. **25**(12), 5892–5904 (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, pp. 886–893 (2005)
15. Enzweiler, M., Gavrila, D.M.: Integrated pedestrian classification and orientation estimation. In: Computer Vision and Pattern Recognition, pp. 982–989 (2010)
16. Ma, W., Huang, D., Lee, N., Kitani, K.M.: Forecasting interactive dynamics of pedestrians with fictitious play. In: Computer Vision and Pattern Recognition, pp. 774–782 (2016)
17. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 73–86. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_6
18. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7575, pp. 201–214. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33765-9_15
19. Ballan, L., Castaldo, F., Alahi, A., Palmieri, F., Savarese, S.: Knowledge transfer for scene-specific motion prediction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 697–713. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_42
20. Rehder, E., Wirth, F., Lauer, M., Stiller, C.: Pedestrian prediction by planning using deep neural networks (2017)
21. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H.S., Chandraker, M.K.: DESIRE: distant future prediction in dynamic scenes with interacting agents. In: Computer Vision and Pattern Recognition, pp. 336–345 (2017)

22. Walker, J., Gupta, A., Hebert, M.: Patch to the future: unsupervised visual prediction. In: Computer Vision and Pattern Recognition, pp. 3302–3309, June 2014
23. Park, H.S., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: Computer Vision and Pattern Recognition, pp. 4697–4705, June 2016
24. Su, S., Hong, J.P., Shi, J., Park, H.S.: Predicting behaviors of basketball players from first person videos. In: Computer Vision and Pattern Recognition, pp. 1502–1510 (2017)
25. Kooij, J.F.P., Schneider, N., Flohr, F., Gavrila, D.M.: Context-based pedestrian path prediction. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 618–633. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_40
26. Schneider, N., Gavrila, D.M.: Pedestrian path prediction with recursive Bayesian filters: a comparative study. In: German Conference on Pattern Recognition, pp. 174–183 (2013)
27. Xie, D., Todorovic, S., Zhu, S.C.: Inferring 'Dark Matter' and 'Dark Energy' from videos. In: International Conference on Computer Vision, pp. 2224–2231, December 2013
28. Yi, S., Li, H., Wang, X.: Pedestrian behavior understanding and prediction with deep neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 263–279. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_16
29. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: human trajectory prediction in crowded spaces. In: Computer Vision and Pattern Recognition, pp. 961–971, June 2016
30. Fernando, T., Denman, S., Sridharan, S., Fookes, C.: Soft + hardwired attention: an LSTM framework for human trajectory prediction and abnormal event detection (2017)
31. Fernando, T., Denman, S., McFadyen, A., Sridharan, S., Fookes, C.: Tree memory networks for modelling long-term temporal dependencies (2017)
32. Lee, N., Kitani, K.M.: Predicting wide receiver trajectories in American football. In: Winter Conference on Applications of Computer Vision, pp. 1–9, March 2016
33. Bokhari, S.Z., Kitani, K.M.: Long-term activity forecasting using first-person vision. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10115, pp. 346–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54193-8_22
34. Rhinehart, N., Kitani, K.M.: First-person activity forecasging with online inverse reinforcement learning (2017)
35. Keller, C.G., Gavrila, D.M.: Will the pedestrian cross? a study on pedestrian path prediction. IEEE Trans. Intell. Transp. Syst. **15**(2), 494–506 (2014)
36. Rehder, E., Kloeden, H.: Goal-directed pedestrian prediction. In: Workshop on International Conference on Computer Vision, pp. 139–147, December 2015
37. Yamaguchi, K., Berg, A.C., Ortiz, L.E., Berg, T.L.: Who are you with and where are you going? In: CVPR 2011, pp. 1345–1352 (2011)
38. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: human trajectory understanding in crowded scenes. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 549–565. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_33
39. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
40. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. Comput. Graph. Forum **26**(3), 655–664 (2007)

41. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: modeling social behavior for multi-target tracking. In: International Conference on Computer Vision, pp. 261–268 (2009)
42. Majecka, B.: Statistical models of pedestrian behaviour in the forum. Ph.D thesis, MSc Dissertation, School of Informatics, University of Edinburgh (2009)
43. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Computer Vision and Pattern Recognition, pp. 3457–3464 (2011)
44. Yi, S., Li, H., Wang, X.: Understanding pedestrian behaviors from stationary crowd groups. In: Computer Vision and Pattern Recognition, pp. 3488–3496 (2015)
45. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Computer Vision and Pattern Recognition, pp. 3354–3361 (2012)