- Object tracking in realistic scenarios is a difficult problem and is an active area of research. Problems : illumination changes, occlusion(hiding), clutter(many objs), camera motion, low contrast(distinguish), specularities(reflection) etc.
- Kaplan-meier metric is used, it is the % of alive patients after treatment; It also deals with censored info where info of a patient is missing for a period. It is used for overall performance.
- Grubbs test : largest abs deviation of $y_i$ from mean considered as outlier. It is used to compare trackers for each vid.
- It was felt that limited no. of videos can't be used to assess robustness. As all algos are quite different, algos were grouped based on experimental performance. Average vid len = 9.2 sec, from diverse events were used as they are hard as model needs to adapt fast.
- Evaluation : GT of first frame is input, evaluated as how well tracker is able to follow target across vid by comparing BB with GT at every 5th frame. Eval is objective as vids have great variety and algos are very diverse.
- Online trackers were used as offline trackers can do forward-backward scanning and find global optimization of path. Pre-trained models aren't used as their performance depends on training data.Vids taken from diverse situations help to analyse strengths….
- Contours based trackers were not evaluated as there are difficulties in case of camera motion and occlusion for initialisation of the contour.
- Previous papers had narrow scope as they focused on only 1 application and eval data was limited (in no. and also application).Some eval metrics were designed without need of GT, but were not able to evaluate all aspects.
- General 3 types of errors : Deviation (from GT), FP and FN. IoU was used with threshold. Dice (2*x*y/(x+y)) was used in previous papers. F1 = mean(2*p*r/(p+r)). P & R are measured at object level as pixel level segmentation was not available.
- Kalman filtering : predicting output of next time step along with its uncertainty based on previous measurements and taking weighted avg with more wt to more certain estimate.
- Particle filter : computing posterior distributions of a markov process using set of particles (samples) which may have partial observations. All particles are given likelihood wt (P(x) of prob distribution) which is used to find new state.
- Hough transform : Feature extraction to find imperfect instances in a certain class. Generalised hough => for detecting arbitrary shapes using template matching.

**Tracking by matching** : Template of target built by model is matched with new frame.

- NCC : candidate windows are sampled uniformly around previous target position and matched using NCC with target template. (No change in target template in this). Window with highest NCC is chosen as new target pos.
- Lucas-Kanade Tracker : Affine transformation (collinearity and ratio of distances preserved) match is applied from target pos of previous frame to windows of new frame by ascent. Affine works by changing scale, rotation and translation.
- Kalman Appearance Tracker : Appearance prediction matching (addresses occlusion) of 20x20 target template with each pixel managed by separate kalman filters, with resized windows. Large diffs are penalised to reduce outlier effect. Window with min diff is selected as new target template.

- Fragments-based Robust Tracking : Fixed Target template is broken into 10x2 patches to address occlusion and pose changes. Windows selected uniformly are matched with target at diff scales (10% up & down) by intensity histogram for each patch by earth movers' dist.Only smallest 25% scores are used for final score of window for robustness.
- Mean Shift Tracking : Matching of initial frame target's RGB color histogram (addresses radical change in shape, texture) with windows using Bhattacharya dist (higher is closer). Mean shift is used to find mode of func (confidence map near prev frame target pos) that maximises BD. (MS finds most similar wrt target)
- Locally Orderless Tracking : Initial BB is segmented into superpixels.Particles are windows made of superpixels. Likelihood of windows is set by parametrized EMD b/n superpixels of target and candidates. Params rep flexibility of target. New target state is likelihood weighted sum of windows using Particle filter with Gaussian wt around prev target pos.

**Tracking by Matching and Extended appearance model** : Extra model for changing the appearance. So matching costs add up as even appearance needs to be matched.

- Incremental Visual Tracking : Eigen images (basis) of target are found by incremental PCA which uses only last few templates. Windows are sampled by Particle filter. Confidence of sample = dist from Eigen image subspace. Window with min score is selected.
- Tracking on the Affine Group : General 2D affine group is used to account for changes other than scale,rotation and translation, from which possible transformations of target are sampled using Gaussian model. It uses same appearance model from IVT.
- Tracking by Sampling Trackers : Maintains extended model of trackers (each made of appearance, motion, state rep and obs models).Target state stores centre,scale and spatial info. Multi scale and pos are considered. IPCA is used with info leakage. Basic trackers made of the 4 components are used to find best target state.

**Tracking Using Matching with Constraints :** Target rep is reduced to a sparse rep which is matched.

- Tracking by Monte Carlo sampling:Focuses on objects whose shapes change drastically by sparse optimization over patch pairs.Target is modeled by sampling a fixed no. of patches described by edge features and color histograms, which are associated with patches from outside. Patches rep by star graph.Best pos in new frame are found by warping new patch to older one for correction. New target pos is found by MAP est. Patches are updated by addn, removal or change of patches.
- Adaptive Coupled-layer Tracking : Fast appearance change using sparse optimization in two local layers constrained by a global layer.Initially target BB divided into patches grid. Locations of patches are estimated by Kalman filter and tuned by affine trans for new frame. Global layer has info on appearance, shape & motion. Color HSV-histograms are used to find appearance likelihood per pixel.Difference b/n velocity of salient points and that of tracker assesses likelihood of motion per pixel. Degree of being inside or outside the convex hull spanned around the patches gives the likelihood of a pixel. Local layers use these 3 likelihoods to update weights of patches and also to calculate overall prob of pixel belonging to target. Global layer slowly updated using stable patches of local lyrs.

- L1-minimization Tracker : sparse optimization by L1 from past appearance. Windows around target are used as bases for sparse rep with non target pixels as alternate bases. Linear combi of windows sampled using Particle filter is formed by L1 minimization. Affine warps of windows are also taken. New target is one with min L1 error.
- L1 Tracker with Occlusion detection : It uses L2 optimization is used to increase speed of above tracker. Candidates below threshold reconstruction error (L2) are only taken for L1 min. To address occlusion, coeffs of alternate bases above a threshold are used. If more than 30% image is occluded, model stops.

**Tracking Using Discriminative Classification** : Classifier that separates target pixels from background pixels which is updated by new samples.

- Foreground-Background Tracker : Incremental Linear discriminant classifier with leakage trained on Gabor texture features of target region and also background. Highest score => new target pos. Color SURF values used instead of intensities.
- Hough-Based Tracking :Classifier with segmentation to address non-rigid and articulated objects. Target is located using Hough forest using initial frame. HOG, f' and f'' are used as features to learn appearance. Pixel with max prob of map is centre of new target. Target is segmented using grabcut algo using sparse pixels near that location to generate new +ve data.
- Super Pixel tracking : Classifier embedded with superpixel clustering by mean-shift using superpixel histograms with a confidence value from overlap of cluster with target BB. Windows of many scales are sampled with Gaussian wt. Superpixel confidence is found using its cluster conf and dist from centre. Window with max sum of conf is new target.
- Multiple Instance learning Tracking : Addresses problem of train data used for Classifier covering background in some +ve ex by training on bags of +ve & -ve ex. Boxes near target BB are used for +ve bag and -ve with boxes at further dist. Windows are uniformly sampled around a circle. Classifier is updated. Haar features are used.
- Tracking, Learning and Detection : Labeled + Unlabeled data for training.Detector learns appearance model from many 2D pattern of target and background.Top 50 detector score windows are selected.Optical flow tracker applies KLT and proposes target window from prev frame. NCC is found for windows and window with max similarity with object model is selected. Then detector model is updated with nearby windows.

**Tracking Using Discriminative Classification with Constraints :** It was observed that classifier correctly classifies pixels which is diff from best target localisation.

- STRuck : It includes labeling process in the learner. S-SVM is used which takes input appearance of sampled window (rep by Haar features) and translation per pixel (label). It is trained on current frame target and windows. Candidate with highest score is new target. Constraint is enforced by learner on S-SVM by keeping target position at maximum score. Locations that violate constraint will become support vectors.

**Analysis**

- Target region : Rectangular or ellipsoid BB. Deliberately some backgd is captured in BB as target-backgd interface is important.
    - Some trackers are contour based which allow max change in appearance but don't generalise.

    - Some use rough segmentation which help in case some part of target is outside. But it requires high robustness as it may introduce false info.
    - Patch based reps allow independent appearance changes. They are more probable in tracking occluded objects, but less in case of rigid or small targets. They don't need precise boundary b/n target and backgd.
    - Multiple BBs also work good as they consider target in different positions rather than optimizing its location near centre.
- Appearance Rep:Assumption that some property is constantly transfered across frames.
    - 2D array may be realistic in case of remote objects. Alternately HSI values or gradients are also used.
    - Histogram removes spatial order and allows target max flexibility and can be successful only for small patches. Spatial info has to captured elsewhere.
    - In feature vectors,at most local order is preserved.Color invariants take care of illumination changes but gives target chance to drift by decreasing gradient mag.
    - In static backg cases it is useful to keep backg intensity rep along with backg intensity prediction. It does not constrain target appearance/motion.
    - Other extra info like other objs for occlusion, illumination condns,zoom,tilt etc.
- Representation of Motion
    - Uniform search is robust as it allows target to move in any dir$^{n}$ but loses target if motion is fast.
    - Probabilistic Gaussian sampling around centre gives more wt to locations near centre and has more bias. It will likely fail if camera is shaking.
    - Kalman filter or optical flow motion prediction reduces search space and helps to track fast moving targets.
    - Implicit motion prediction as in MLT/KLT is by maximizing some func. But it assumes that motion is small compared to appearance change.
    - In Detection + Tracking, many candidates are matched with target proposed by optical flow which is robust to fast motion. Recovery from drift is also useful.
    - Prediction of pos, scale, or motion may be useful for extended occlusions when the object is likely to continue to change and no update of the model is possible.
- Method : Finding best target location and state.
    - Best match : Direct gradient ascent on score/probabilities. Particle filter is used to cover more area when comp power is more.But it has strong assumptions about appearance which is problematic if intensity/albedo change. Probabilistic is helpful for occluded objs & low contrast but if obj is detectable, Pb = direct match.
    - Subspace sampling is used with extended model of appearance which keeps targets's appearance across time(helpful for long-term tracking during occlusion).
    - Constrained optim works good if constraints are reliable. As constraints have to come from first few frames, it can be applied to pre-trained tracking.
    - Tracking using classification is semi-supervised as after 1st frame only unlabeled data is there. So it has to learn from 1st few frames for stability. It is robust as it allows very large feature sets. It assumes that target has unique features. Problem with MIT is wrongly labeled -ve instances may cause drifting.

- Model Updating
    - Fixed models don't add false info and may perform well for short sequences. Changing template can work for slow motion and long-term occlusions. So changing target should be done carefully to prevent insertion of false info.
    - Updating patches are complex decisions based on limited data => difficult.
- Evaluation : Plotting sorted outcome (eval) of all videos is survival curve (50% IoU). For comparing trackers for each video, Grubbs' test is used to look for outstanding trackers (99 % confidence and F>0.5). Survival curves of any trackers are compared by log rank statistics.Square of log rank statistic is calculated which approx follows Chi-square distn.
    - To visualise effectiveness of eval metrics, they are plotted against each other. F1, ATA, OTA had high correlation(pairwise with F) so F-score is used. Deviation is used to measure tightness of tracking by threshold. It has low correlation with F and OTP.
    - STR, FBT, TST, TLD and L1O are the best, and each is from diff category of trackers. 4 fold eval on stochastic trackers produces nearly same curves => data is sufficient,stable.
    - With an ideal combi of all trackers, a region is found which is not yet possible to track in the curve (approx 10%). Ideal combi is much better => improve scope.
    - Deviation is used to measure acc. NCC, KLT and L1O (direct match) are best. TLD, MST and SPT are worst as their motion is limited.But diffs are small so F.
- Trackers performance was analysed based on type of difficulty with avg F-scores normalised across all aspects for each tracker.
- Illumination Conditions : It depends on whether tracker can diff b/n target and shadow. There is significant diff b/n avg F-score of target-only models and target+backg models. FBT is best, and trackers working good in this include backg.
- Changes in the Target's Surface Cover : It also benefits from including backg by significant amt. FBT is best again.
- Specularities : L1T performs outstanding as it used single pixel templates which prevent effect of specular pixels. But L1O detects specularity as occlusion and stops tracking.
- Shape and Size : Performance is lowest with shape change compared to all aspects. L1T is best as it places shape constraints and is flexible. Plotting mean F-score against binned target size shows that MST and ACT depend on target size. IVT,SPT and LOT also depend but less. MST may get stuck in local optima for small targets. IVT and LOT may benefit from larger target as they have max no. of params. SPT and LOT use superpixels which may not work good for small targets.Discriminative work good for all.
- Target's Motion:Trackers have a lot to improve on smoothness and coherence of motion. But they can work well with extreme appearance changes without apparent motion and full-short occlusions.
- Clutter and Confusion : Performance is good but becomes complicated if illumination changes. They track objs which are difficult for humans also.
- Occlusion : Motion of target w.r.t camera is imp. Under no/little motion, performance is good for partial as well as short-full occlusions.Top 5: STR, **FBT**, TST, and **TLD** and

L1T. Short (30%) occlusion with relative motion are also tracked but not with full occlusion.

- Camera Motion : TLD adapts to cam motion as detector finds for target throughout img and tracker finds among nearer pos. IVT performs relatively good (Gaussian motion).
- Zoom : IVT is best. 8 of 19 trackers perform well for gradual size change. Mechanism to tackle zooming is beneficial in general. KLT also performs well due to warping.
- Length of the Video : STR, FBT, NCC and TLD perform good.
- Overall even performance : STR (except zoom), TST (except long vids due to many params => false info), L1O (except smoothness), TLD (except surface cover), NCC (except cam motion and zoom).
    - TMC performs poorly maybe due to its complex method which highly depends on quality of initial BB. Patches are sampled such that their amt depends on intensity variations. If initial BB contains backg => much false info added.
    - TAG loses target faster than models without affine capacity, as it fails to estimate transformation correctly due to many params that have to be estimated reliably.
    - In ACT, small error in the global layer leads to larger error in target localisation.
- Accuracy of the Initial Bounding Box : By shifting BBs to right by 20%, robustness is evaluated. Overall loss in performance is less, with IVT, TAG and MST highly affected. STR, HBT and TST are very robust.
- Ideal combi shows with large margin from best trackers shows that each tracker is better at some circumstances.
- BBs have proved to be useful and are part of most successful trackers. Inclusion of backg adds robustness to illumination changes and appearance changes.
- FBT performs best for illumination changes and surface change which indicates some value of SURF features. But it doesn't know how to deal with shape change and cam motion.
- Using Gaussian favors current pos more, which helps in case of no/little apparent motion and also short occlusions.
- More complex models (update) introduce more degrees of freedom for error and perform worst in long videos.
- STR's overall solid performance is attributed to use of S-SVM that optimises displacement from target pos. It only fails on zooming changes due to lack of capacity.
- TLD also performs good overall except on illumination changes and appearance changes. TST also except long vids due to complexity. L1O improves on L1T on aspects other than occlusion. Generally if trackers perform good in 1 field, other fields go down.
- Student's  t-test was performed to check superiority of group of trackers in some fields.
- Trackers which solve (label) drifting, small selections of large sets of features, and low-complexity update mechanisms without losing too much on the current achievements will progress.

- To evaluate performance correctly for hard occlusion cases, a new dataset is formed which covers categories of hard occlusion. They show that trackers can't be evaluated by avg score as their performances vary wildly across diff categories.
- The toughest and rarest of the category vids have to be tracked to enable deployment. It was felt that occlusion, in-plane rotation, and out-of-plane rotation are under-rep in the standard datasets. No distribution exists in parts of the object that are occluded.
- Correlation Filters: It uses circular corr in which all shifted variants of target are included that creates a powerful appearance model with limited data. CF is fast with Fourier trans. MOSSE tracker enabled use of multi-dim features, improved robustness to scale change and deformations, mitigating boundary effects, and the use of deep convolutional filters. ECO reduces complexity of CF tracker by using a factorised CNN to increase speed.
- SiamFC finds a similarity func (using FC siamese network) offline which is used to find obj. SiamRPN uses RPN to generate proposals for more accurate BBs. ATOM improves on Siam by using a target estimation and target classification model.
- To address occlusion, Siam nets have structured dropout, ROT overcomes target model decay. An exp strategy analyses occurrence of occlusion using spatiotemp context info, also motion constraints and target reference are helpful. Another uses a layer-based approach with backg occluding layers.
- Many datasets categorize occlusion into partial, full and out-of-frame. They are simple as they have little apparent motion of target and for short durations. LaSOT has more challenging cases but they are limited.
- HOB consists of 20 long sequences annotated with BB at every 15th frame. General cases of occlusion occur with strong motion and scale-changes. Types of occlusion :
    - Full out of frame occlusion (FOC) : for long periods and target may enter from diff location compared to exit location.
    - Feature occlusion (FO) : specific features are occluded.
    - Occlusion by transparent object (OCT) : visible but appearance changed.
    - Occlusion by similar object (OCS)
- ECO, SiamFC, SiamRPN++, ATOM and DiMP are evaluated. 3 variants of SiamRPN++ were evaluated (Resnet50, shallow Alexnet backbones and a long-tem update strategy). ATOM and DiMP use a memory model to keep track of appearance changes over time.
- One pass evaluation (OPE) is done to measure precision, succes-rate, AUC, and least-subsequence-metric for all 20 HOB videos.
    - Precision : Center localization error (Euclidean dist b/n centres) for thresholded true +ves. As it is sensitive to resolution, all vids in HOB have same resolution.
    - Success Rate : It measures position acc and BB shape & size acc. BB overlap is done by IoU over thresholded +ves. IoU = 1 if tracker correctly tells obj is absent.
    - Least Subsequence Metric : It is the ratio of longest successfully tracked subsequence to length of vid. Successful => p% of subseq is tracked with IoU > t for each fr.It is biased towards extremely long & short vids but here all are same.

- Top 20 hard vids from LaSOT are evaluated for baseline. On an average, performance on HOB was worse than LaSOT. SiamRPN++ with Resnet50 backbone was best on HOB and beats long-term one by small margin inspite of the fact that It was made for long term tracking. This implies that It might have stuck to a wrong target for long time.
- On LaSOT DiMP is best on AUC and LSM metrics with SiamRPN++ It best on precision and comparable scores on other 2 metrics. Others are significantly low. DiMP underperforms both Siams on HOB but HOB is not biased towards Siams, So DiMP overfits on its train data (LaSOT). Also DiMP does frequent updates even in occlusion which adds bias.
- FOC : SiamRPN++ It is best due to re-initialisation of target loss. SiamRPN++ r50 is good due to its rich features at diff scales. DiMP loses due to updates which causes appearance model decay.
- OCS : Siam alex is best. Lt's re-initialisation does not add up so it is slightly lower.
- FO : r50 and It are best.
- OCT : DiMP is best with alex next.
- Top performing trackers can vary b/n diff types of occlusion drastically.